Miss rate neglect in legal evidence

CHRISTIAN DAHLMAN†
Law Faculty, Lund University, Box 117, 22100 Lund, Sweden

FRANK ZENKER
Department of Philosophy, Lund University, Box 192, 22100 Lund, Sweden; SAS, Institute of Philosophy, Klemensova 19, 813 64 Bratislava, Slovakia; and Konstanz University, Philosophy, Postfach D9, 78457 Konstanz, Germany

AND
FARHAN SARWAR
Department of Psychology, Lund University, Sweden

[Received on 2 May 2016; revised on 4 July 2016; accepted on 26 September 2016]

Research on probabilistic reasoning has discovered several systematic errors, among which base rate neglect and the fallacy of the transposed conditional have featured prominently. This article introduces the term miss rate neglect to capture the systematic failure to properly account for false positives, i.e. the probability of evidence (E) given the hypothesis (H) is false, P(E|~H). Miss rate neglect occurs when decision makers (i) completely disregard the miss rate; (ii) underestimate the importance of differences in the miss rate, or (iii) overlook circumstances that affect the miss rate. We explain the relevance of miss rate neglect for legal decision making, review extant literature, present new experimental work that empirically validates options (ii) and (iii), and propose experimental variations that future research may pursue.

Keywords: base rate neglect; heuristics and biases; inverse fallacy; legal decision making; miss rate neglect; shoe print experiment; two witnesses experiment.

1. Introduction

Research on probabilistic reasoning has led to the discovery of several systematic errors, which have been described as fallacies. The failure to take account of prior probabilities was discovered by Daniel Kahneman and Amos Tversky,† and is known as ‘base rate neglect’ or the ‘base rate fallacy’. The incorrect inference that a conditional probability is equal to the inverted conditional probability, P(A|B) = P(B|A), was discovered in medical diagnosis by David Eddy,‡ investigated in legal reasoning by William Thompson and Edward Schumann,§ and has since been discussed by scholars as the

† Corresponding author. Email: Christian.Dahlman@jur.lu.se
‘inverse fallacy’ or the ‘fallacy of the transposed conditional’. Some scholars view ‘base rate neglect’ as more important than the ‘inverse fallacy’, and claim that people commit the inverse fallacy due to base rate neglect, while other scholars hold the opposite view, and claim that the ‘inverse fallacy’ causes ‘base rate neglect’.

In this article, we introduce the term ‘miss rate neglect’ as a description of a systematic failure to take proper account of false positives, i.e. the probability of the evidence (E) given that the hypothesis (H) is false, $P(E|\neg H)$. The experiments that we report on show that subjects pay insufficient attention to the miss rate, and underestimate the importance of differences in the miss rate. In our view, this is an important error that should be highlighted as a systematic error in probabilistic thinking. Miss rate neglect can lead to an assessment where the relevance of evidence is seriously over-rated, or seriously under-rated.

We first provide an overview of base rates, hit rates and miss rates (Section 2); then discuss base rate neglect (Section 3), the inverse fallacy (Section 4), as well as miss rate neglect (Section 5); report two experiments on ‘miss rate neglect’ in the assessment of legal evidence (Sections 6 and 7); and discuss them in relation to previous experiments on ‘base rate neglect’ and the ‘inverse fallacy’. Our conclusions are summarized in Section 8.

2. Base rate, hit rate and miss rate

A decision-maker who is presented with additional evidence (E) for some hypothesis (H), must reassess the probability of H and calculate the probability of the hypotheses given the new evidence, $P(H|E)$. How evidence affects this probability is generally captured by ‘Bayes Rule’. In the ‘odds version’ ‘Bayes Rule’ can be stated as follows:

$$\frac{P(H)}{P(\neg H)} \times \frac{P(E|H)}{P(E|\neg H)} = \frac{P(H|E)}{P(\neg H|E)}$$

$P(H)$ is the ‘prior probability’, i.e. the probability of the hypothesis before the evidence has been considered. $P(\neg H)$ is the prior probability that the hypothesis is false, before the evidence has been considered. $P(H)$ and $P(\neg H)$ are additive, in the sense that $P(H) + P(\neg H) = 1$. In a legal case, where the hypothesis states that the defendant is guilty and the evidence consists in the testimony of an eyewitness, $P(H)$ is the probability that the defendant is guilty before the testimony has been taken into account. The ratio of the relevant prior probabilities, $P(H)/P(\neg H)$, is referred to as the ‘prior odds’, i.e. the odds that the hypothesis is true before the evidence has been considered. For instance, if $P(H) = 20\%$, then $P(H)/P(\neg H) = 0.20/0.80 = 1/4$.

Some assessments use a statistical ‘base rate’ as the prior probability. As we shall see, assessments that underestimate the importance of a statistical base rate commit the fallacy known as ‘base rate neglect’.

$P(E|H)$ is the probability of the evidence if the hypothesis is true. In the legal example above, it is the probability that the eyewitness ‘would’ give the testimony if the defendant ‘were’ guilty. $P(E|H)$ is referred to as the ‘hit rate’ or the ‘probability of a true positive’. Analogously, $P(E|\neg H)$ is the

probability of the evidence if the hypothesis is false. In the legal example, it denotes the probability that the eyewitness would give the testimony if the defendant were innocent. The terms ‘miss rate’ or ‘probability of a false positive’ refer to \( P(E|\sim H) \). It is crucial to notice that the hit rate and the miss rate are not additive. \( P(E|H) \) and \( P(E|\sim H) \) need ‘not necessarily’ add to 1. The ratio between the hit rate and the miss rate, \( P(E|H)/P(E|\sim H) \), is called the ‘likelihood ratio’. If, e.g. the hit rate is 96% and the miss rate is 8%, the likelihood ratio can be calculated as \( 0.96/0.08 = 1.2 \).

\( P(H|E) \) is the ‘posterior probability’, i.e. the probability of the hypothesis after the evidence has been considered. In our example, it is the probability that the defendant is guilty after the eyewitness testimony has been considered. Analogously, \( P(\sim H|E) \) denotes the probability that the hypothesis is false after the evidence has been considered. In our example, \( P(\sim H|E) \) denotes the probability that the defendant is innocent after the testimony has been considered. \( P(H|E) \) and \( P(\sim H|E) \) are additive just like \( P(H) \) and \( P(\sim H) \), so \( P(H|E) + P(\sim H|E) = 1 \). When expressed as a ratio, \( P(H|E)/P(\sim H|E) \) is called the ‘posterior odds’, i.e. the odds that the hypothesis is true after the evidence has been considered.

‘Bayes Rule’ dictates that the prior odds multiplied by the likelihood ratio equals the posterior odds. As an example using the numbers above, prior odds of 1:4 and a likelihood ratio of 12 yields a posterior odds of 3:1, since \( 1/4 \times 12 = 3 \). A posterior odds of 3:1 means that the posterior probability, \( P(H|E) \), is 75%, since \( 0.75/0.25 = 3 \). In other words, the evidence has raised the probability of the hypothesis from 20% to 75%.

In legal theory, the likelihood ratio is sometimes described as the ‘force’ of the evidence.\(^6\) If the likelihood ratio \( > 1 \), the evidence increases the posterior probability of the hypothesis. When the likelihood ratio \( < 1 \), the evidence decreases the posterior probability. In the event that the likelihood ratio \( = 1 \), the evidence is irrelevant for the hypothesis, and does not affect the posterior probability.

3. Base rate neglect

In 1973, Daniel Kahneman and Amos Tversky conducted experimental work on probabilistic reasoning to test whether participants took proper account of base rates.\(^7\) In what became known as the ‘Tom W. Experiment’, participants were divided in three groups: BASE-RATE, SIMILARITY and PREDICTION.

The BASE-RATE group was asked to consider all first-year college students in the USA, and to assess their distribution over nine fields of specialization: business, computer science, engineering, humanities, law, library science, medicine, physical science and social science. Participants assessed humanities to be the most common specialization among college students; the mean value for humanities was 20%, and the mean value for computer science only came to 7%.

The SIMILARITY group was given the following description of a student called ‘Tom W’.

Tom W. is of high intelligence, although lacking in true creativity. He has a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by some corny puns and by flashes of imagination of the sci-fi type. He has a strong drive for competence. He


\(^7\) Kahneman & Tversky supra note 1, p. 49–50.
seems to have little feel and little sympathy for other people and does not enjoy interacting with others. Self-centered, he nonetheless has a deep moral sense.

Participants in the SIMILARITY group were asked to assess Tom’s similarity with the ‘typical’ student in each of the above nine fields of study, upon which participants assessed Tom as being most similar to the typical student in computer science. The mean similarity rank for computer science was 2.1, while humanities only received a mean similarity rank of 7.2.

The PREDICTION group was given the same description of Tom W., and was asked to assess the probability that Tom W. is studying each of the nine fields of specialization. Assessments in the PREDICTION group strongly correlated with results from the SIMILARITY group: participants judged it to be most probable that Tom studies computer science (the mean prediction rank for computer science was 2.6 and the mean prediction rank for humanities was 7.6).

Kahneman and Tversky concluded that participants in the PREDICTION group made their probability assessment on the basis of similarity, and ignored the base rate. If they had taken account of the 7% base rate for computer science, and the 20% base rate for humanities, they should have realized that the probability of a randomly picked student being enrolled in computer science is affected not only by his similarity to a typical student of computer science, but also by the overall percentage of students in this field of specialization. Hence, if participants had taken proper account of the relevant base rates, the prediction rank would not have correlated so well with the similarity rank. In particular, the difference between the ranks for computer science and humanities should have been far smaller in the PREDICTION group than in the SIMILARITY group.

The broad failure of those in the PREDICTION group to properly take account of the base rate has come to be known as ‘base rate neglect’. It should be noted that the ranking of similarity between Tom W. and the typical student in each specialization involves the same kind of question that arises in the assessment of the likelihood ratio, where the hypothesis (H) says that Tom is studying a certain specialization and the evidence (E) consists in personal information about Tom. Put into probabilistic terms, the hit rate, \( P(E|H) \), is the probability that a student with that specialization would have that personality, and the miss rate, \( P(E|\neg H) \), is the probability that a student in some other specialization would have that personality. The assessment of the likelihood ratio, \( P(E|H)/P(E|\neg H) \), therefore involves the same kind of question as the ranking of similarity, namely: ‘Is it more common that a student has a certain personality if he studies computer science than if he studies some other specialization?’

The ‘Tom W. Experiment’ has been widely discussed in the literature on probabilistic thinking, and it has spawned a number of follow-up studies that have tested base rate neglect in other contexts. An experiment on base rate neglect in a legal context was conducted in 2003 by Mark Schweizer. The experiment had 173 Swiss judges as participants, and we will refer to it as the ‘Breathalyzer Experiment’. Participants were asked to consider an automobile driver who is stopped by the police for a routine alcohol test (random selection), exhales into a Breathalyzer and tests positive. Judges were asked to assess the probability that the driver is in fact driving under the influence of alcohol, given the following information:

---

8 Ibid p. 51.
One in a hundred drivers is driving under the influence. If someone is under the influence of alcohol there is a 95% probability that the Breathalyzer will show a positive result. Unfortunately, there is a 5% probability that a person who is not under the influence will get a positive result.

Participants were asked to assess the probability that a driver is in fact under the influence of alcohol provided that the Breathalyzer shows a positive result. They were given four alternative answer ranges: 0–25, 26–50, 51–75 or 76–100%, where the correct answer is the interval 0–25%.

To see this, consider that the probability of the hypothesis (H) that the driver is under the influence of alcohol given the evidence (E) that the Breathalyzer shows a positive result, \( P(H|E) \), can be calculated using 'Bayes Rule'. Since the prior probability, \( P(H) \), is 1% (i.e. 1 in 100 drivers), the hit rate, \( P(E|H) \), is 95%, and the miss rate, \( P(E|\sim H) \), is 5%, it follows that the posterior probability, \( P(H|E) \), is approximately 16% \( \left( \frac{0.01}{0.99} \times 0.95 / 0.05 \approx 0.16/0.84 \right) \). Only 9% of the judges arrived at the correct answer. Most judges, 68%, responded that the probability falls in the 76–100% range. Schweizer, therefore, concluded that a majority of the judges did not take proper account of the base rate.10

The research on base rate neglect has been criticized by some scholars who claim that people are not as inclined to neglect base rates in real-world contexts as the ‘Tom W. Experiment’ and the ‘Breathalyzer Experiment’ suggest.11 Critics maintain that experiments demonstrating base rate neglect have been intentionally designed to distract participants from the base rate, and that results would have been different had the experiments been set up in a way that made the participants more aware of the base rate. Gerd Gigerenzer claims that ‘base rate neglect’ disappears completely if participants are made aware of the underlying frequencies that determine the base rate.12 Gigerenzer conducted an experiment where the set-up required of each participant to draw a description of a person from an urn, in contrast to the set-up of the ‘Tom W. Experiment’ where participants were just given the description. The participants in Gigerenzer’s experiment did not neglect the base rate.13

4. The inverse fallacy

Probabilistic reasoning that commits ‘the inverse fallacy’ confuses the conditional probability \( P(A|B) \) with the inverted conditional probability \( P(B|A) \), to the effect that these probabilities are treated as equal, i.e. \( P(A|B) = P(B|A) \). In ‘Bayes Rule’, this error can happen in two ways: the hit rate, \( P(E|H) \), can be mistaken for the posterior probability that the hypothesis is true, \( P(H|E) \), and the miss rate, \( P(E|\sim H) \), can be mistaken for the posterior probability that the hypothesis is false, \( P(\sim H|E) \). To distinguish these errors from each other, we refer to the former as ‘hit rate inversion’ and to the latter as ‘miss rate inversion’.

David Eddy is credited for having discovered ‘hit rate inversion’ in medical diagnosis when reviewing literature on mammography.14 Eddy found several examples where the probability that a

10 Ibid, p. 162.
13 Ibid s. 13.
patient has cancer given a positive X-ray report, \( P(\text{H|E}) \), had mistakenly been equated with the probability that an X-ray reports positively given that the patient has cancer, \( P(\text{E|H}) \). Hit rate inversion may thus mistakenly convert a high hit rate into a high posterior probability that the hypothesis is true, disregarding the base rate as well as the miss rate. If, e.g. the hit rate is 95%, ‘hit rate inversion’ will result in the belief that the posterior probability is 95%, which could be completely of the mark. The posterior probability can be low even if the hit rate is high—namely, when the base rate is low, or the miss rate is high. The probabilities in the ‘Breathalyzer Experiment’ above provide for an illustration. As we saw there, the hit rate was 95%, but the posterior probability only came to 16%, owing to a base rate of 1% and a miss rate of 5%. In fact, the main result of the ‘Breathalyzer Experiment’ could be explained as an effect of ‘hit rate inversion’.15 Most participants (68%) placed the posterior probability in the 76–100% range, which is exactly what you ‘should’ believe if you equate the posterior probability with the hit rate. Since ‘hit rate inversion’ may thus cause ‘base rate neglect’, some scholars have argued that ‘base rate neglect’ should be viewed as an effect of the ‘inverse fallacy’, rather than a fallacy in its own right.16

Chris Guthrie, Jeffrey Rachlinski and Andrew Wistrich have conducted an experiment on ‘hit rate inversion’ in legal evidence, based on the English case Byrne v. Boadle.17 The participants were 159 judges from federal courts in the USA. Byrne v. Boadle is a tort case from 1863, where the plaintiff had been hit by a barrel of flour that fell from a second-story loft in a warehouse belonging to the defendant. The participants in the Byrne v. Boadle Experiment were given the following information.18

The defendant’s employees are not sure how the barrel broke loose and fell, but they agree that either the barrel was negligently secured or the rope was faulty. Government safety inspectors conducted an investigation of the warehouse and determined that in this warehouse: (1) when barrels are negligently secured, there is a 90% chance that they will break loose; (2) when barrels are safely secured, they break loose only 1% of the time; (3) workers negligently secure barrels only 1 in 1000 times.

The participants were then asked to assess the probability that the barrel fell due to negligence, and were given four alternative ranges: 0–25, 26–50, 51–75 or 76–100%. The correct answer, here, is 0–25%. As above, the probability of the hypothesis (H) that the barrel was negligently secured given the evidence (E) that the barrel fell, \( P(\text{H|E}) \), can be calculated using ‘Bayes Rule’. Since the prior probability, \( P(\text{H}) \), is 0.1%, the hit rate, \( P(\text{E|H}) \), 90% and the miss rate, \( P(\text{E|\sim H}) \), 1%, it follows that the posterior probability, \( P(\text{H|E}) \), is approximately 8% (0.001/0.999 * 0.90/0.01 ≈ 0.08/0.92).

Of all participants, 41% arrived at the correct answer, 0–25%. Another, 9% answered 26–50%, 10% answered 51–75% and 40% stated 76–100%.19 As a possible explanation, Guthrie and colleagues submit that judges who responded 76–100% committed ‘hit rate inversion’.20

---

15 This possibility is mentioned by Schweizer in his discussion of the results from the Breathalyzer Experiment. See Schweizer supra note 10, p. 162.
19 Ibid p. 809.
20 Ibid p. 810.
The set-up in the Byrne v. Boadle Experiment is structurally identical to Schweizer’s ‘Breathalyzer Experiment’. But responses falling in the 76–100% interval are subsumed under entirely different explanations. Guthrie and colleagues explain their data as an ‘inverse fallacy’ ('hit rate inversion'), whereas Schweizer explains his data as instances of ‘base rate neglect’. For some scholars, the important finding in these experiments is that people do not always take proper account of base rates. To others, the experiments rather demonstrate that a conditional probability may be mistaken for an inverted conditional probability, and ‘base rate neglect’ is merely a side-effect of this error.21

‘Miss rate inversion’ in legal evidence was discovered by William Thompson and Edward Schumann when they studied the argumentation of prosecutors in criminal trials.22 As an example, Thompson and Schumann report a case where the perpetrator had a blood type found in 10% of the population that matched the blood type of the defendant. The prosecutor went on to interpret this match as a 90% probability that the defendant was guilty, \(P(H|E)\)—suggesting a line of reasoning that confuses the miss rate, \(P(E|\neg H)\), with the posterior probability that the hypothesis is false, \(P(\neg H|E)\). The fact that the blood type is found in 10% of the population means that the miss rate is 10%, i.e. the probability that the defendant has the relevant blood type (E) if the defendant is innocent (\(\neg H\)) is 10%, but this does not mean that the posterior probability that defendant is innocent, \(P(\neg H|E)\), is 10%. The prosecutor incorrectly inferred that the posterior probability of innocence is 10%, and, therefore, incorrectly concluded that the posterior probability of guilt, \(P(H|E)\), is 90%. Thompson and Schumann referred to the error as the ‘prosecutor’s fallacy’, and this term has caught on in legal literature. They also conducted an experiment where participants who acted as mock jurors in a criminal case heard a prosecutor present an argument that committed the prosecutor’s fallacy.23 Of them, 29% assessed this argument as ‘correct’.

Notice that ‘miss rate inversion disregards’ the base rate, just like ‘hit rate inversion’.

5. Miss rate neglect

It is not the purpose of this article to take sides in the debate between scholars who view ‘base rate neglect’ as the principal problem and scholars treating the ‘inverse fallacy’ as the principal problem. Rather, in our view, previous research on errors in probabilistic thinking has paid insufficient attention to the failure of taking proper account of the miss rate. We think that ‘miss rate neglect’ should be treated as an important fallacy in its own right. There are some experiments, where it has been observed that people are insufficiently attentive to miss rates,24 and there are some scholars who have discussed the ‘neglect of alternative hypotheses’,25 but these studies have so far not been duly appreciated in the literature on ‘base rate neglect’ and ‘inverse fallacy’.

It should be noted that results of the ‘Breathalyzer Experiment’ and other similar experiments that are presented in support of ‘base rate neglect’ could just as well be presented in support of ‘miss rate neglect’. After all, the low posterior probability in the ‘Breathalyzer Experiment’ (16%) is not a consequence of a low base rate only (1%), but just as much a consequence of a comparatively high

21 Villejoubert & Mandel supra note 5, p. 172, with further references.
22 Thompson & Schumann supra note 3, p. 170.
23 Ibid p. 177–178.
miss rate (5%). If the miss rate had been lower—say, 0.5% instead of 5%—the posterior probability would have been higher (66% instead of 16%). Therefore, the result of the experiments can just as well be presented in support of ‘miss rate neglect’. The participants who incorrectly assessed the posterior probability to be >25% failed to appreciate that the miss rate was not sufficiently low to bring the posterior probability above 25%. And experiments such as the *Byrne v. Boadle* Experiment that demonstrate ‘hit rate inversion’ could just as well be presented in support of miss rate neglect.

An assessment of probability that commits ‘miss rate neglect’ fails to take proper account of how the posterior probability, \( P(H|E) \), is affected by the miss rate, \( P(E|\sim H) \). In the most extreme cases of ‘miss rate neglect’, the assessment does not take account of the miss rate at all, as in ‘hit rate inversion’. In less extreme cases, the miss rate is not neglected completely, but its importance for the posterior probability is not recognized properly. For, as we saw, it makes a vast difference to the value of the posterior probability whether the miss rate is 5 or 0.5%. Probabilistic thinking that fails to appreciate such differences also commits ‘miss rate neglect’.

Insufficient attention to the miss rate may also result in the incorrect assessment that some circumstance that affects the miss rate is irrelevant for the calculus. Imagine, e.g. a burglary case, where one piece of evidence consists in a shoe print that shows that the burglar has the same shoe size as the defendant. This print constitutes ‘stronger’ evidence against the defendant if the shoe size is more unusual, i.e. if the size is ‘less typical’ for a given population segment, e.g. 46 (EUR)/12½ (U.S.), in comparison to a more common size, e.g. 44 (EUR)/10½ (U.S.), since the miss rate is lower when the size is more unusual. A judge who fails to appreciate this, and, therefore, thinks that the shoe size is irrelevant for the probability that the defendant is guilty, commits ‘miss rate neglect’. This mistake should be distinguished from the fallacy known as ‘confirmation bias’, where evidence against a hypothesis is overlooked.\(^{26}\) The difference between ‘confirmation bias’ and ‘miss rate neglect’ is that a judge who acts under ‘confirmation bias’ overlooks the evidence because he is committed to the hypothesis. The judge believes that the hypothesis is true, or wants to believe that it is true. By contrast, ‘miss rate neglect’ entails no such commitment.

In sum, ‘miss rate neglect’ can be committed in three well-distinguishable ways by:

1. completely disregarding the miss rate
2. underestimating the importance of differences in the miss rate
3. overlooking circumstances that affect the miss rate

To further investigate ‘miss rate neglect’ in probabilistic reasoning, experiments targeting each of these three variants should be conducted. The first variant has already been demonstrated in the *Byrne v. Boadle* Experiment and further experiments on ‘hit rate inversion’. As far as the authors know, the second variant has not been experimentally tested, while the third variant has been observed in some experiments, but has not been tested in a systematic way.

Ruth Beyth-Marom and Baruch Fischhoff have conducted an experiment on ‘pseudo-diagnosticity’, where they observed that participants overlooked circumstances that affect the miss rate. We refer to it as the ‘Bear’s Club Experiment’. Participants were asked to evaluate the hypothesis that ‘Mr. Maxwell is a university professor’ based on the evidence that he is a member of the Bear’s Club. Participants were also asked to assess how relevant certain questions would be to this investigation. One of these questions concerned the probability that Mr Maxwell ‘would’ be a member of the Bear’s Club if he

---

were not a university professor. This question obviously explores the miss rate, and is therefore relevant, but 65% of all participants assessed it as irrelevant to the investigation. This can be interpreted as a demonstration of ‘miss rate neglect’.

We conducted two further experiments on ‘miss rate neglect’. The ‘Two Witnesses Experiment’ tests the second way of committing ‘miss rate neglect’ (underestimating the importance of differences in the miss rate), and the ‘Shoe Print Experiment’ tests the third variant (overlooking circumstances that affect the miss rate).

6. The Two Witnesses Experiment

In the Two Witnesses Experiment, 211 participants were recruited from three categories: 28 professional judges working at municipal courts in Sweden, 75 Swedish barristers and 108 law students from a university in Sweden. Participants were presented with the following scenario:

In a criminal trial two eye witnesses, A and B, are testifying for the prosecution. A and B are independent witnesses. They have not communicated with each other. The prosecutor has called A and B to the stand to establish that the culprit fled from the scene of the crime in a blue car. When the prosecutor asks about the colour of the car, A says that it was difficult to see in the dark, but the car appeared to be blue. B is asked the same question, and gives the same answer. The defence attorney argues that the witnesses could be mistaken about the colour. In the dark, a black or green car can be mistaken for blue. To test if the witnesses could be mistaken about the colour, a reconstruction of the event is conducted on the same location under the same light conditions. A and B observe 100 cars of different colour passing by, and are asked to indicate the colour of each car.

The result of the test is displayed in the following Tables 1 and 2.

Participants were asked to assess how strongly these testimonies supported the prosecutor’s claim that the observed car was blue, and to compare the support provided by both witness A with the support

<table>
<thead>
<tr>
<th>Table 1. Result for blue cars</th>
</tr>
</thead>
<tbody>
<tr>
<td>The witness testifies that the car is blue (correct) (%)</td>
</tr>
<tr>
<td>Witness A</td>
</tr>
<tr>
<td>Witness B</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. Result for non-blue cars</th>
</tr>
</thead>
<tbody>
<tr>
<td>The witness testifies that the car is blue (incorrect) (%)</td>
</tr>
<tr>
<td>Witness A</td>
</tr>
<tr>
<td>Witness B</td>
</tr>
</tbody>
</table>

27 Beyth-Marom & Fischhoff supra note 24, p. 1190.
given by witness B. Instructions asked participants to indicate (a) whether A’s testimony provides stronger support to the prosecutor’s claim than B’s testimony, (b) whether their testimonies each provide equally strong support to the prosecutor’s claim or (c) whether B’s testimony provides stronger support than A’s claim. Participants who answered that one testimony gave stronger support than the other were also asked if they perceived the difference as ‘slight’ or ‘substantial’.

The correct answer is: (c) B’s testimony gives stronger support to the prosecutor’s claim than A’s testimony. The likelihood ratio, \[ P(E|H)/P(E|\neg H) \], is 0.90/0.20 = 4.5 for A’s testimony, but 0.80/0.10 = 8 for B’s testimony. This means that the evidentiary force of B’s testimony is about twice as strong as the evidentiary force of A’s testimony. When the probability that the car is blue is updated on the basis of B’s testimony, the prior odds that the car is blue, \[ P(H)/P(\neg H) \], are multiplied by a factor of 8. When the probability that the car is blue is updated on the basis of A’s testimony, the prior odds are only multiplied by a factor of 4.5.

Our hypothesis stated that most participants would not fully or not at all appreciate the importance of this difference in the miss rate, and assess the probabilities incorrectly. Results from the experiment are displayed in Table 3.

Results (Table 3) are in line with our hypothesis. While 39% of judges gave the correct answer, a sizable majority (61%) of the judges arrived at an incorrect assessment of the probabilities. A goodness-of-fit test qualifies these results as statistically significant deviations from a random model (\( \chi^2(2, N = 28) \approx 12.35, p < 0.01, \text{Cohen’s } w = 0.66 \)). Slightly fewer correct answers were obtained from barristers (35%) and law students (36%). These results were strongly significant against a random model (barristers: \( \chi^2(2, N = 75) \approx 27.44, p < 0.0001, \text{Cohen’s } w = 0.60 \); law students: \( \chi^2(2, N = 108) \approx 17.39, p < 0.001, \text{Cohen’s } w = 0.40 \)).

Most participants (53%) came to the incorrect assessment that the testimonies of A and B provide equally strong support to the prosecutor’s claim (answer ‘b’). Interviews conducted after the experiment suggest that these participants arrived at their assessment because the difference between A and B ‘appeared symmetrical’. Witness A is superior in correctly identifying blue cars (Table 1), while witness B appears to be equally superior in correctly identifying non-blue cars (Table 2). But this line of reasoning is incorrect, as a difference between 80% and 90% in the hit rate is less important than a difference between 10% and 20% in the miss rate. If the hit rate climbs from 80% to 90%, the likelihood ratio increases by 12.5%. If the miss rate drops from 20% to 10%, the likelihood ratio increases by 50%. Participants who said that A and B provide equally strong support to the prosecutor’s claim thus committed ‘miss rate neglect’ by underestimating the importance of differences in the

<table>
<thead>
<tr>
<th></th>
<th>Judges (%)</th>
<th>Barristers (%)</th>
<th>Law students (%)</th>
<th>All (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) A’s testimony gives stronger support to the prosecutor’s claim than B’s testimony.</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>b) A’s testimony and B’s testimony give equally strong support to the prosecutor’s claim.</td>
<td>57</td>
<td>57</td>
<td>48</td>
<td>53</td>
</tr>
<tr>
<td>c) B’s testimony gives stronger support to the prosecutor’s claim than A’s testimony.</td>
<td>39</td>
<td>35</td>
<td>36</td>
<td>36</td>
</tr>
</tbody>
</table>

TABLE 3. Results from the ‘Two Witnesses Experiment’ on miss rate neglect; proportion of answers given by judges (\( N = 28 \)), barristers (\( N = 75 \)) and law students (\( N = 108 \))
miss rate (the ‘second variant’ of ‘miss rate neglect’ discussed in Section 5). Notice that this error cannot be interpreted as ‘base rate neglect’ or ‘inverse fallacy’. Since participants in the experiment were only asked to assess the force of the evidence, not the posterior probability, there was no base rate to neglect, and no posterior probability to confuse with the hit rate or miss rate. In order to explain the error, ‘miss rate neglect’ must be established as a fallacy ‘sui generis’.

The result from the ‘Two Witnesses Experiment’ calls for follow-up studies. It would, e.g. be interesting to conduct an experimental variation of the experiment, where witnesses provide ‘conflicting’ testimonies about the colour of the car (A testifies that the car was blue; B testifies that it was green). In this scenario, it makes a bigger difference which testimony carries the stronger evidentiary force.

7. The Shoe Print Experiment

Participants in the ‘Shoe Print Experiment’ were 123 law students from a university in Sweden, who were given the following scenario:

MK is standing trial for burglary. According to the prosecutor, MK broke into a warehouse and stole two BMW motorcycles. MK denies the charges. During the trial, GA testifies as a witness for the prosecution. GA says that he passed the warehouse one night when he was out walking his dog and observed a person rolling a motorcycle into the back of a van. He hid in the bushes, and observed the burglar from a distance of approximately 40 meters. It was dark, but the van was parked under a street light. GA identifies the defendant MK as the man who stole the motorcycles. In addition to this evidence, the police found ten shoe prints from the burglar at the scene of the crime. The shoe prints have the same size as MK’s shoes, and come from Nike trainers size 44 (EUR)/10½ (US). MK claims that he has nothing to do with the burglary. He has been previously convicted for stealing a car, and this is admitted as character evidence against him.

Participants were asked to consider five situations in which circumstances of the case had been changed in ways described below (items a–e), and to assess whether a given change strengthens the case against MK or not. Participants were asked to indicate all of the changes that strengthen the evidence against MK.

(a) GA knows MK personally. They are members of the same athlete club.
(b) The police found twenty shoe prints from the burglar at the scene of the crime.
(c) MK has been previously convicted for stealing a motorcycle.
(d) The shoe prints have the same size as MK’s shoes, and come from Nike trainers size 46 (EUR)/12½ (US).
(e) GA observed the burglar from a distance of approximately 20 meters.

The purpose of the experiment was to test how participants assessed the change described in item d. The other changes served as filler items. The correct answer for item d is that the change in shoe size strengthens the evidence against MK (see Section 6), because 46 (EUR)/12½ (U.S.) is more unusual than 44 (EUR)/10½ (U.S.). The change reduces the miss rate, and thereby increases the likelihood ratio as well as the posterior probability. Our hypothesis stated that most participants would overlook the relevance of item d, and commit miss rate neglect by assessing the change in shoe size as not strengthening the evidence. Results were in line with this hypothesis. Of all participants, 30% gave
the correct answer and 70% incorrectly assessed the change in shoe size as not strengthening the case for the prosecution. A goodness-of-fit test qualifies this result as a strongly significant deviation from a random model ($X^2(1, \ N=123) \approx 19.52, \ p < 0.0001, \ Cohen\'s \ w = 0.40$).

It can be assumed that the participants knew that 46 (EUR)/12½ (U.S.) is a more unusual shoe size than 44 (EUR)/10½ (U.S.). This means that the participants knew that it is less likely that MK would have shoe size 46 (EUR)/12½ (U.S.) than 44 (EUR)/10½ (U.S.), if he is an innocent person mistakenly identified as the burglar. The failure to notice that this is relevant for the force of the evidence, and strengthens the case against MK, shows a lack of attention to circumstances that affect the miss rate (i.e. the ‘third variant’ of ‘miss rate neglect’ discussed in Section 5).

The results broadly replicate findings from the ‘Bear Club Experiment’ (see Section 5), where 65% of all participants overlooked a circumstance affecting the miss rate. In our experiment, 70% overlooked such a circumstance.

The results from the ‘Shoe Print Experiment’ call for follow-up studies. It would, e.g. be interesting to conduct a variation of the experiment, where the difference in shoe size is more extreme, e.g. size 44 (EUR)/10½ (U.S.) versus size 48 (EUR)/14 (U.S.).

8. Conclusions

This article has introduced the term ‘miss rate neglect’ for a systematic failure in probabilistic reasoning. On our view, ‘miss rate neglect’ deserves more attention, and should be treated as a fallacy in its own right.

As we saw, an evidentiary assessment that commits ‘miss rate neglect’ fails to take proper account of the probability of the evidence if the hypothesis is false, $P(E|\sim H)$. We also saw that ‘miss rate neglect’ can be committed in three different ways.

(1) completely disregarding the miss rate
(2) underestimating the importance of differences in the miss rate
(3) overlooking circumstances that affect the miss rate

We reported the results from two experiments on ‘miss rate neglect’ in legal evidence. In the ‘Two Witnesses Experiment’ a clear majority of participants underestimated the importance of differences in the miss rate. In the ‘Shoe Print Experiment’, an equally sizable majority of participants overlooked circumstances positively or negatively affecting the miss rate.

We invite others to attempt to replicate these results. Other than replication, future research should, of course, investigate de-biasing methods, i.e. interventions that assist judges in taking proper account of the miss rate.

Acknowledgements

Research financed by grant from Ragnar Söderbergs stiftelse.