Erling Eide*

Two tests of the base rate neglect among law students

Abstract

The present paper presents an experiment designed to test law students' capacity to apply probabilistic reasoning in determining the likelihood of a defendant's guilt. As expected, most students are victims of base rate neglect, possibly related to a representative heuristic, the statement of an eyewitness. The results might be taken as a warning against the use of probabilistic thinking in courts, the intuition of judges could easily lead to wrong decisions. It is argued, however, that the problem does not lie with the formal modelling of probabilities, but in our intuition when dealing with uncertain evidence. In order to avoid miscarriage of justice future judges should be acquainted with Bayesien reasoning.[1]

JEL NO: K40, C11, C91

*Erling Eide, Dept. of Private Law, University of Oslo, Box 6706, 0130 Oslo, Norway.
erling.eide@jus.uio.no

1 INTRODUCTION

[1] A preliminary version of this paper, A Test of the Base Rate Neglect among Law Students", was presented at the 28th Annual Conference of the European Association of Law and Economics, in Hamburg, September 22-24, 2011.

1.1 PURPOSE

Are judges capable of using probabilistic evidence in a satisfactory manner? The booming literature on behavioural science shows that people tend not to take uncertain information correctly into account when making decisions.[2] In addition, with the progress of forensic science and empirical studies on stochastic events, probabilistic evidence is becoming more easily obtainable.[3] One may wonder whether judges are just as likely as other people to misuse such evidence, with miscarriage of justice as a consequence.

Judicial decision making may be considered as a problem of decision making under uncertainty (Lando 2006). When the uncertainty is characterised by probabilities, decision theory, in particular the Bayes'Theorem[4], provides rules for how correct decisions can be reached. Correct conclusions might nevertheless be hard to obtain. One obstacle is the requirement of computational competence. Another is the shortage of rationality revealed in studied in behavioural science. Both problems are widespread (Stanovich and West 2000). The present paper focuses on the latter difficulty.

Although the literature on how people misunderstand probabilistic information is abundant, empirical studies of how judges treat probabilistic evidence are few.[5] Information that makes it possible to determine how judges reason more in detail about probabilistic evidence is seldom available in court cases. Moreover, judges may not be very interested in participating in laboratory tests that might reveal incompetence.[6] In the present study, I have substituted master law students for judges. The study might possibly reveal how well prepared law students are for eventually dealing with probabilistic evidence as judges.

One particular problem is the "base rate neglect". The phrase describes the tendency for people to disregard or underestimate certain relevant statistical information, the base rates.[7] If judges reason like other people, base rates as pieces of evidence are neglected, or at least not given appropriate weight. The purpose of the present paper is to test whether the problem of base rate neglect is present for future judges. If it is, one may as a consequence hypothesize that miscarriage of justice occurs.

Judges have two main tasks: establish facts and make verdicts. Mistakes might obtain in both activities. In order to avoid miscarriage of justice, the uncertainty of facts has to be correctly taken into account. In constructing various models of cognition, psychologists usually distinguish between intuitive processes and deliberative processes (Kahneman & Frederick, 2002, at p. 51). Both processes, sometimes combined, are present in court cases. To the extent that evidence is uncertain,

---

2 See, in particular Kahneman and Tversky (1972) and the great number of studies in behavioural economics that have appeared thereafter.

3 See e.g. Jobling and Gill (2004) on DNA and Brink et al. (2007) on fingerprints and handwriting.

4 This theorem follows from the multiplicative rule of probability, which holds that the joint probability of two events, say B and W_B, equals the product of the conditional probability of one of the events, given the second event and the probability of the second event. In mathematical notation:

$$P(W_B \cap B) = P(W_B|B)P(B)$$
$$P(W_B \cap B) = P(B|W_B)P(W_B)$$
$$\Rightarrow P(B|W_B) = \frac{P(W_B|B)P(B)}{P(W_B)}$$

5 A notable exception is Strnad (2007).

6 I requested the permission of the Norwegian Association of Judges to carry out at their annual meeting a simple test similar to those presented in this paper, but received a negative answer explaining that they, as a principle, did not accept such proposals.

7 A "base rate" may be defined as the relative frequency with which an event occurs or an attribute is present in a population (Ginossar and Trope 1987). Hence the base rate for men in a country might be higher than 0.51, whereas the base rate for men in the army of such counties might be less than 0,05.

decision theory provides procedures that should be applied in the deliberative processes. To reach a verdict all pieces of evidence have to be assessed in order to decide whether a relevant criterion, such as “beyond reasonable doubt” in criminal cases, or “more probable than not” in civil cases, has been met. Decision theory hardly solves the whole decision process; its domain is the deliberative aspects.

In the literature there is disagreement about to which extent scientifically based deliberations are common in court cases. Koehler (1996) has tested judges’ general reasoning skills as well as their decision-making skills in legal contexts. He concludes that judges, like other people, commonly make judgments intuitively, rather than reflectively, both generally and in legal contexts. [8] A similar conclusion is reached by Guthrie et al. (2007). The present study does not argue against these empirical findings on how judges in fact reason. The question is, with more probabilistic evidence now available, are future judges fit to deal with the uncertainties? The focus is on cognition of uncertainty and not on what judges do or have to do in dealing with legal questions. The topic is the applicability of the normative model of decision making.

The result of a test carried out by Guthrie et al. (2001) is an indication of the seriousness of the problem. They asked a group of federal magistrate judges to evaluate the probabilistic evidence of the classic English case, *Byrne v. Boadle*.[9] Only 40 per cent of the judges steered clear of the base rate neglect. In a later article the authors remark (Guthrie et al. 2007, p. 24) that compared with results obtained in studies of other people (in particular Cassells et. al. 1978), ”the overall relative performance of judges was admirable”. This evaluation must be a meagre consolation for defendants that eventually were judged by the 60 per cent of the judges that neglected base rates.

2 THE TEST

The purpose of the test is to investigate the occurrence of base rate neglect among law students. Over the years, I have carried out more or less the same test with a number of groups of master students in law.[10] I here present the results of two tests carried out in 2010 and 2011, test No. 1 and test No. 2, respectively. The problem to be solved is similar to the “blue and green cab problem” discussed in the literature. Koehler (1996, p. 7) has argued that the results of these studies are not as conclusive as the authors think, because the probabilities presented to the respondents have been ambiguous. In the present study some effort is taken in order to

[8] Koehler (1996) argues, however, that in the literature there is a common misunderstanding that experiments reveal a general base rate neglect. He argues (p. 4) that more often than not base rates are not neglected, but often are given less weight than what is appropriate. On the lawyer-engineer problem he concludes that “the data do not provide strong support for the conclusion that base rates are ignored or even “largely ignored” (Kahneman and Tversky 1973, p. 242). Instead, the data indicate that, at least under some conditions, base rates influence probability judgments considerably, and sometimes to a degree that satisfies an abstract normative standard.”

[9] The judges were presented following with the summary of the evidence: “The plaintiff was passing by a warehouse owned by the defendant when he was struck by a barrel, resulting in severe injuries. At the time, the barrel was in the final stages of being hoisted from the ground and loaded into the warehouse. 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 authors then asked: “Given these facts, how likely is it that the barrel that hit the plaintiff fell due to the negligence of one of the workers?” The judges were asked to answer by choosing one of the following four probability ranges: 0-25%, 26-50%, 51-75%, or 75-100%. Only about 40 % selected the (correct) low range.

[10] The experiment is in the tradition from Finkelstein and Faily (1970) and Lempert (1977).

formulate the problem such that the probabilities stand out as unquestionable facts. Moreover, the base rate is given an exceptionally low value in order for the students not to overlook it.

At the end of one of my lectures in law and economics for master law students in 2010 (before speaking about probabilities in procedure) I asked the students to solve the following problem, presented on a questionnaire:[11]

> Suppose a car accident has taken place on an island without road connection to the mainland. The car has disappeared. Assume that in order to assess the colour of the car, only two pieces of evidence are available:
>
> > 1. There is only one (ordinarily trustworthy) eyewitness to the accident. According to the eyewitness the car was blue. The judge is informed by a psychologist that ordinarily trustworthy eyewitnesses mistakenly say that in 5 % of such circumstances the car was blue when it was not (5 % false positives). Furthermore they say that the car was blue in all cases when it in fact was blue.
> >
> > 2. On the island 0.1 % of the cars are blue.
>
> NOTE: The judge believes that these pieces of information are correct.
>
> *The judge wishes to determine the probability that the car was blue. Which probability (in per cent) should the judge arrive at?*

One might argue that because students are not accustomed to deal with probabilities, a test like the present one is of little relevance to court decisions. Moreover, one might argue that evidence could be uncertain, but not probabilistic. In order to explore this criticism, the text of test No. 2 was slightly changed compared to the one of test No. 1. The following paragraph was substituted for assumption 2 of Test No. 1

> On this island there are very few blue cars. In a speed control carried out before the accident took place, in which photographs were taken of all cars, only one in about 1000 were blue.

Moreover, the two assumptions were interchanged, i.e. that in test No. 2 the assumption about the prevalence of blue cars was presented before the assumption of what the eyewitness reported.

At variance with some base rate studies in the literature the base rate in the present study was assumed to be very low. The reason for this choice was to have the students not to ignore this assumption. Furthermore, in order to avoid having the students questioning the relevance of the two assumptions (pieces of evidence), it was underscored in the questionnaire that the judge believed that the assumptions were correct. Moreover, the students were given the opportunity to comment on the problem at the end of the questionnaire. None of them took the opportunity to discuss the assumptions.

3 RESULTS

The number of students present in test No. 1 was 55. Of these, 48 delivered questionnaires where marks indicated their choice of probability.

[11] See Appendix 1 for a detailed description of the procedure of the test and Appendix B for the questionnaire. The questionnaire is translated from Norwegian by the author.

Table 1 presents the distribution of the answers. Clearly, the answers of a majority of students are concentrated around the probability that the statement of an eyewitness is correct. The frequency of blue cars on the island is neglected by a great majority of the students. Only about 10 per cent have taken the base rate properly into account. Some of the comments these students have given on their questionnaires indicate that they had some knowledge about the content of the Bayes' Theorem.

The Bayes' theorem gives the statistically correct answer.[12] Let B be the event that the car was blue, NB the event that the car was not blue, W_B the event that the eyewitness states that the car was blue, and W_{NB} the event that the car was not blue. By assumption, $P(W_B)=P(W_B|B)P(B)+P(W_B|NB)P(NB)$. Substituting into the Bayes' formula $P(B)=0.001$, $P(W_B|B)=1$, $P(W_{NB})=0.999$ and $P(W_B|NB)=0.05$, one obtains

$$P(B|W_B) = \frac{P(W_B|B)P(B)}{P(W_B)} = \frac{1 \cdot 0.001}{1 \cdot 0.001 + 0.05 \cdot 0.999} = \frac{0.001}{0.05095} \cong 0.02$$

Table 1. Test No. 1. Number of answers in various

intervals of probability that the car was blue

Probability in per cent	Number of answers
0-3	5
4-10	1
11-20	0
21-40	2
41-60	2
61-80	2
81-90	0
91-96	32
97-100	4

An intuitive solution to the problem is easily obtained if the probabilities are interpreted as frequencies. Suppose 1001 accidents identical to the one at hand have taken place. Because the prevalence of blue cars on the island (the base rate) is 0.1 percent, statistically only about one of the 1001 accidents is caused by a blue car. The eyewitness, however, will (because of the 5 % false positive) say that the car was blue not only in this accident, but also in 50 of the other accidents. Thus, the eyewitness is correct only in one of 50 cases. In probabilities: 1/51 = 0,01999 or about 2 %.

Only 10 % of the students answered that the probability of the car being blue was less than 3 %. The distribution of answers is remarkably similar to a test carried out among physicians by Casscells et al.

12 The Bayes' Theorem is in various fields routinely employed as a normative model for aggregating base rate and other probabilistic information, see e.g. Slovic and Lichtenstein (1971).

(1978). In this study “...20 house officers, 20 fourth-year medical students and 20 attending physicians, selected in 67 consecutive hallway encounters at four Harvard Medical School teaching hospitals, [were asked] the following question: “If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5 per cent, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person’s symptoms or signs?” (p. 999) As the authors explain, only one of 1000 people studied will, on average, have the disease, and 5 per cent of the others, or roughly 50 persons will, on the average, yield (falsely) positive results. The chance that any one positive result represents a person with the disease would be slightly less than 2 per cent. Eleven of 60 participants, or 18 per cent, gave the correct answer. The most common answer, given by 27, was 95 per cent, and the average of all answers was about 56 percent, a 30-fold overestimation of disease likelihood.

The number of students participating in test No. 2 was 79 (all students present answered the questionnaire). The results are summarized in Table 2.

Table 2. Test No. 2. Number of answers in various

intervals of probability that the car was blue

Probability in per cent	Number of answers
0-3	4
4-10	5
11-20	3
21-40	1
41-60	8
61-80	2
81-90	11
91-96	37
97-100	3

One observes that the tendency of wrong answers is about the same as in test No. 1. There are, however, some differences. In test No.2 the proportion of correct answers is somewhat smaller than that of test No. 1 (4 out of 79 as against 5 out of 55). But given the small number of these correct answers, the difference is insignificant. Of some interest is the fact that the proportion of answers mirroring the 95% that a statement of an eyewitness is reliable is substantially lower in test No. 2 than in test No. 1 (37 out of 79, or 26 %, as against 32 out of 55, or 58 %). Moreover, in test No. 2 there are 30 answers in the interval 4-90 %, but only 7 in the same interval in test No. 1. It seems that the change from probabilistic evidence (0.1 % blue cars) to just more uncertain evidence (about 1 in 1000 blue cars) has produced less precise opinions among the students. The difference could also be the result of the change in the order that the two main assumptions are presented.

Taken together, only a small minority in the two tests answer correctly, and a majority seem to count completely or mainly on the reliability of the eyewitness. One might guess that the eyewitness serves as a representative heuristic as explained in behavioural science.

4 Discussion

The main result of both tests do not deviate substantially from a number of base rate neglect studies found in the literature (see, however Koehler 1996). Master students in law are neglecting base rates at least to the same extent as other people. As future judges, they seem to be unprepared to carry out the scientific deliberations required in order to avoid miscarriage of justice.

What could be the relevance of the tests for real court cases? Two main problems must be addressed. Is the base rate fallacy as common in the court room as in experiments? How common is it that evidence depends on base rates, presented as probabilities or eventually in more vague terms?

Koehler (1996) argues that the proponents of the representative heuristic have overstated the extent to which people actually neglect base rates. He maintains that although in a number of practical situations people underestimate the *weight* that should be given to base rates, such rates usually to some extent influence judgements. People do not *neglect* base rates, but they sometimes place *too little weight* on them. The results of test No. 2 might be considered as a kind of substantiation of his claim. There, the spread of answers indicates that the base rate has been given a modest weight by some of the students. The great majority of them are, however, far of the mark.

Koehler also argues that the degree of underweighting depends on the circumstances and problems to be solved. Although this seems reasonable, one might question the relevance of the argument when tests as simple as the present ones reveal substantial underweighting of the base rate. Would not more complex circumstances make it even more difficult to take base rates correctly into account?

It is certainly true, as Koehler states, that verdicts are never based on probabilistic evidence alone. Such pieces of evidence are usually included in a broader set. There seems to be scant information about the proportion of evidence characterised by base rates or on more vague information of similar kind. Although judges informally argue that such evidence is rare, no systematic information exists for Norwegian court cases. However, infrequency of base rates does not solve the problem in cases where probabilistic evidence has a decisive influence. One might of course presume that such cases are rare. Test No. 2, however, indicates that the problem is present also when the base rate is not presented as a probability. Therefore, regardless of how base rates are expressed or formulated, to the extent that such rates are part of the evidence, one might expect wrong verdicts.

Given the results of the present tests and of other studies in the literature, one might be tempted to conclude that probabilistic reasoning should be avoided by judges. Some of them, perhaps a majority, will be lead astray, and wrong decisions will be taken.

This conclusion is vulnerable to the argument that the avoidance of probabilistic reasoning will not prevent the judges' intuition to lead them astray. The problem of the base rate neglect might be present also when uncertain evidence is not available as probabilities. The result of test No. 2 underscores this point. The more vague presentation of the low prevalence of blue cars did not increase the number of correct answers. Vague uncertain evidence does not seem to be easier to handle than corresponding probabilistic evidence.

Eliminating all intuition from judicial decision making is both impossible and undesirable because it is an essential part of how the human brain functions. Intuition, however, might cause wrong decisions not only because people have to rely on it, but because they rely on it when it is inappropriate to do so.

If any normative conclusion can be drawn from the test, it is that judges should be trained in Bayes' Theory, not only in order to avoid problems where the base rate is clearly expressed as probabilities or proportions, but also when such rates are presented in more vague terms.[13] More generally, they should also be trained in aggregating probabilities, both base rates and other probabilistic information.

5 CONCLUSION

The study demonstrates that the base rate neglect was present among almost all master law students that participated in a laboratory experiment. In deciding whether a car causing an accident was blue or not blue a great majority ignored the stated information about the low prevalence of blue cars. Instead, they fully trusted an eyewitness saying that the car was blue, although they were informed that the witness could be mistaken in some such cases.

The tests indicate that there is hardly any reason to believe that the students' confusion about how to evaluate base rates and uncertainty is restricted to probabilistic evidence. Only a very small minority of the participants in the tests reasoned according to the Bayes' Theorem. This result demonstrates that the future judges (as well as acting judges) should be vaccinated against the base rate neglect by some sort of education in statistics. Because people apparently find is easier to reason about frequencies than about probabilities, the education should include an explanation of how probabilities could be interpreted as frequencies.

[13] This conclusion is in line what has been recommended by other researchers in various fields including forensic science (Evett 1986) and law (Fienberg and Schervish 1986, Finkelstein 1987, Kaye 1989). I agree with the final statement by Strnad (2007, p. 292): "Yes, legal empiricists should go Bayesien".

References

Brink, Axel, Lambert Schoemaker and Marius Bulacu (2007): Towards explainable writer verification and identification using vantage writers," 2 icdar, 824-828,

Casscells, Ward, Arno Schoenberger and Thomas B. Graboys (1978): Interpretation by physicians of clinical laboratory results, 299 *New England Journal of Medicine*, 18, 999–1001.

Evett, I. W. (1986): A Bayesian Approach to the Problem of Interpreting Glass Evidence in Forensic Science Casework. 26 *Journal of the Forensic Science Society*, 3-18.

Fienberg, Stephen E. (1986): Gatecrashers, blue buses, and the Bayesian representation of legal evidence. 66 Boston University Law Review, 693-699.

Finkelstein, M. and W. Fairly (1970): A Baysian Approach to identification of evidence, 83 *Harvard Law review*, 489-517.

Ginossar, Zvi and Yaacov Trope (1987): Problem solving in judgment under uncertainty. 52 *Journal of Personality and Social Psychology*, 464-474.

Guthrie, Chris, Jeffrey J. Rachlinski and Andrew J. Wistrich (2007): Blinking on the bench: how judges decide cases. 93 *Cornell Law Review*, 1-43.

Jobling, Mark A. and Peter Gill (2004): Encoded evidence: DNA in forensic analysis. 5 *Nature Reviews, Genetics*, (October), 739-751.

Daniel Kahneman, and Shane Frederick (2002): Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In: *Heuristics and biases: the psychology of intuitive judgement*, ed. by Thomas Gilovich, Dale Griffin and Daniel Kahneman. Camebridge University Press, 49-81.

Kahneman, Daniel and Amos Tversky (1972): Subjective probability: a judgement of representativeness. 3 *Cognitive Psychology*, 430-454.

Kahneman, Daniel and Amos Tversky (1973): On the psychology of prediction. 80 *Psychological Review*, 237-251.

Koehler, Jonathan J. (1996): The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. 19 *Behavioral and Brain Sciences*, 1–53.

Lando, Henrik (2006): A derivation of probabilities of correct and wrongful conviction in a criminal trial. 2 *Review of Law and Economics*, 371-379.

Lempert, R. O. (1977): 75 Modeling relevance. *Michigan Law Review*, 1021-1057.

Sanchirico, Chris William (2010): Models of evidence: survey and assessment. Research Paper No. 10-28, University of Pennsylvania, Institute of Law and Economics.

Slovic, Paul and Sarah Lichtenstein (1971): Comparison of Bayesian and regression approaches to the study of information processing judgement. 6 *Organizational Behavior and Human Performance*, 649-744.

Sonnemans, Joep and Frans van Dijk (2010): Errors in judicial decisions: experimental results. Forthcoming in *Journal of Law, Economics and Organization*.

Stanovich, Keith E. and Richard F. West (2000): Individual differences in reasoning: implications for the rationality debate? 23 *Behavioral and Brain Sciences*, 645-726.

Strnad, Jeff (2007): Should Legal Empirisists Go Bayesian? 9 *American Law and Economics Review*, 195-303.

Appendix 1: Details on the test procedure

At the end of one of my lectures 15 minutes were allocated to the test. First, I gave the students the following oral information (translated from Norwegian)

"Before the break, I would like you to answer a question related to the question of characterizing uncertain evidence by probabilities. The test is an element in our endeavour to link together the various subjects you are studying this term. [These subjects include procedure.] After having distributed a sheet of paper presenting the problem, I shall give you some more information about the background for the problem."

*

Then, the questionnaire (Appendix 2) was distributed.

*

"Most evidence is uncertain, even evidence that often is considered to be certain, like DNA evidence and finger prints.

In some cases the uncertainty of evidences are presented as probabilities. Examples are "P-pilledommen II" and "Landåssaken" [the Landås case] in which a nursing cadet was accused of having murdered several patients. In the legal literature there is disagreement about to which extent evidentiary uncertainty should be presented and evaluated as probabilities. Such presentations occur, however, and because a concept like probability preponderance is commonly employed, it is important for judges and juries to evaluate probabilities.

Today's questionnaire is made with the purpose of obtaining information about how difficult it might or might not be to evaluate uncertain evidence.

In order for the test to be worthwhile, you should not speak to each other before delivering your answer. The answer should, as far as possible, be given under the same conditions as at written exams. Everybody should deliver the sheet of paper, even if the question is not answered, in order to determine the proportion of participants that have answered.

At the end of the sheet of paper there is a place where you may add your comments to your answer (or lack of answer).

I will present the results next week, when procedure is the topic of my lecture.

You should deliver your answer before the break – i.e. within 10 minutes.

Appendix 2

Evaluation of uncertain evidence

Suppose a car accident has taken place on an island without road connection to the mainland. The car has disappeared.

To simplify, assume that in order to assess the colour of the car, only two pieces of evidence are available:

1. There is only one (ordinarily trustworthy) eyewitness to the accident. According to the eyewitness the car was blue.

   The judge (the jury) is informed by a psychologist that ordinarily trustworthy eyewitnesses make mistakes in the sense that they in 5 % of such situations say the car was blue when it was not (5 % false positives). Furthermore they say that the car was blue in all cases that it in fact was blue.

2. On the island 0.1 per cent (1 per thousand) of the cars is blue.

NOTE: The judge is assumed to believe that this information is correct.

*

The judge wishes to determine the probability that the car was blue.

Which probability (in per cent) should the judge attain?

Put a mark against chosen interval

Probability in per cent	
0-3	
4-10	
11-20	
21-40	
41-60	
61-80	
81-90	
91-96	
97-100	

Comment, e.g. if no mark is given: