

## LAWYERS' LIKELIHOODS

Peter Hawkins, College of Law, UK  
Anne Hawkins, RSS Centre for Statistical Education,  
The University of Nottingham, UK

*This paper explores whether lawyers can understand and apply the criminal burden of proof, and whether they are able to make the sort of likelihood estimates that the expert witness suggested would help the jurors in the case of R v Adams. The English Court of Appeal expressly rejected the use of mathematical formulae to determine guilt or innocence. With this kind of restraining judgment, it is important that lawyers should have a thorough and reliable grasp of probability. Results suggest that their areas of misunderstanding include not only problems in making coherent estimates of likelihood or probability, and then drawing logical inferences, but also difficulties with percentages and ratios, and even with the criminal and civil burdens of proof.*

### INTRODUCTION

This study was undertaken in the wake of the controversial English legal decisions in the case of *R v Adams* (1996, 1997), where the Court of Appeal expressly rejected the use of Bayes' theorem on the grounds that '... the attempt to determine guilt or innocence on the basis of a mathematical formula, applied to each separate piece of evidence, is simply inappropriate to the jury's task'. In the second appeal, the Court of Appeal went even further, stating that expert evidence should not be admitted that might encourage jurors to attach mathematical probability values to items of qualitative evidence.

This paper builds on an earlier study by the authors (1998), that demonstrated a number of classical statistical and probabilistic misconceptions to which lawyers are susceptible. The current study investigated the specific problems that lawyers have in handling stochastic information like that presented in the *Adams* case. In particular, it explored lawyers' understanding of the criminal burden of proof (i.e. beyond reasonable doubt); whether they can derive valid (albeit subjective) assessments of probability; whether they can deal effectively with disparate items of quantitative and qualitative evidence; and the extent to which their judgments are based on *absolute* (bigger or smaller than) comparisons or *relative* (number of times bigger or smaller than) comparisons. The emphasis was on the degree to which the respondents made *logical* judgments (given their beliefs about the accused's likelihood of guilt and their standards of proof) and the circumstances under which they might revise their thinking. There was no external criterion by which to determine what was the 'correct' verdict in this study.

## RESPONDENTS, METHODOLOGY, AND THE RESEARCH INSTRUMENT

The respondents (218 in all) had just begun a Legal Practice Course. They were graduates either in law or in another subject followed by a conversion course to law. All had received at least one year's training in law, including courses covering legal method and criminal law. The present study required each respondent to complete a paper and pencil Schedule entitled *Interpreting Courtroom Evidence*. No conferring was allowed, and the respondents were told that their responses would be made anonymously.

The Schedule asked each respondent to state his or her gender, in case males and females responded differently to the rape scenario. Question A(i) asked them to express the criminal burden of proof as a percentage. An extra question, A(ii), was put to 84 respondents, i.e. how many *times more likely* must it seem that a defendant is guilty than innocent before they would convict him or her. Next followed a case-study incorporating the major features of the *Adams* case, i.e. the quantitative prosecution DNA evidence (the chance of the DNA being from some other man was one in 2 million) *versus* the qualitative non-DNA evidence (the possibility that the attacker was not a local man, plus the non-identification of the defendant, his alibi, and alibi support evidence). Question B then asked all the respondents to express as a percentage the likelihood that the accused was guilty, and Question C required them to opt for a verdict of 'guilty' or 'not guilty'.

The second page of the Schedule attempted to make the lawyers rethink the case in the way proposed by Professor Donnelly (the expert statistical witness for the defence in the *Adams* case). This involved producing percentage likelihoods for evaluating the non-DNA evidence. After doing this, respondents were again instructed to reach an overall verdict of 'guilty' or 'not guilty'.

## BELIEFS ABOUT THE CRIMINAL BURDEN OF PROOF - PAGE 1 OF THE SCHEDULE

The responses to Question A demonstrated that some of the law students were not reliably handling the criminal burden of proof. Although the median values of the responses for both males and females were 90%, the responses showed great variability, ranging from 100% down to 50%, this last figure possibly indicating a confusion with the *civil* burden of proof (balance of probabilities). Six females, but only one of the males, made this mistake. At the other extreme, a greater percentage of the females specified more demanding standards of proof, e.g. 100%, or absolute certainty (18% of females c.f.

9.75% males). Also, more of the females specified standards in the very high nineties (i.e. 97% - 99%). The males' stated levels for the burden of proof were generally lower, which *might* suggest that they were less worried about making false conviction errors.

The question about how many times more likely a person must seem to be guilty than innocent for the respondent to convict him or her provided an alternative way for stating a standard of proof. There was no evidence of a direct correspondence between the respondents' expressed percentages for the criminal burden of proof and the standard implied by their 'times' judgments, apart from some erroneous equating of percentages and 'times' (particularly noticeable in 13% of the respondents who gave 90%-100% as the criminal burden of proof). A comparison of these 'times' judgments with their beliefs in likelihood of guilt (converted to 'times' terms) was not a good predictor of the verdicts chosen. Although significant at the 0.05 level, it only successfully predicted 51% of the verdicts. There was also a great deal of variability in the 'times' responses associated with the initial decision of 'guilty', ranging from 1.4 times to 100 times, and considerable overlap with those associated with 'not guilty'. In general, there was much to suggest that respondents had problems in handling comparisons based on *relative* (times) assessments.

#### BELIEFS ABOUT THE GUILT OF THE ACCUSED

By way of contrast, most respondents chose a verdict in response to Question C that reflected a coherent absolute (as opposed to a relative) judgment. The absolute comparison variable (their perceived likelihood of the accused's guilt minus their own standard of proof) was significant at the 0.0001 level. It successfully predicted 85% of stated verdicts, whereas the *simple* belief in likelihood of guilt (although significant) could only predict 54% of the verdicts, further confirming the notion that some kind of subjective (intuitive) *comparison* was being used by the respondents. However, 40 respondents (19.5%) responded 'guilty' although logic would have indicated a 'not guilty' verdict. Of these, 42.5% originally gave 100% as the criminal burden of proof. This suggests that some respondents might believe in an *ideal* standard of proof, but be willing to interpret it flexibly in a real situation. Eggleston (1983) highlighted a number of factors that might influence judgments about the burden of proof. The present authors suggest that possibly the judgment of likelihood of guilt is really a *subjective indication of the respondent's confidence in his or her own decision*, rather than a direct estimate of the

accused's likelihood of guilt. This would merit further research.

#### WEIGHING THE SPECIFIC ITEMS OF EVIDENCE - PAGE 2 OF THE SCHEDULE

Page 1 of the Schedule required little more of the respondents than a judge would expect of a jury in a real case, namely to hear the evidence and to decide on a verdict which satisfies the criminal burden of proof. Page 2 (albeit prompting more specific numerical estimates to be associated with the qualitative evidence) also asked the respondents to do what judges say that jurors already have to do, namely to weigh up all the items of evidence against one another in order to reach their verdict.

In the *Adams* case, the jurors were shown how to make *subjective* likelihood estimates for each item of non-DNA evidence (likelihood that the actual rapist was local; and then, first given guilt and second given innocence, two likelihood judgments for each of - failure to identify, alibi and alibi support). They were then shown how to turn these conditional likelihood estimates into likelihood *ratios* that could be accumulated to give an overall estimate of the likelihood of the accused being not guilty. This, they were told, could be directly compared with the prosecution's DNA probability evidence if they chose to do so. There was no suggestion of the witness trying to answer the 'ultimate question'.

In the present project, the respondents were also encouraged to derive the above likelihood estimates. The researchers wanted to know whether or not the respondents would be able to make such estimates at all. In fact, the respondents did appear to be capable of recognising and producing the different conditional probabilities of guilt or innocence as separate entities. Only a minority of respondents gave equal estimates for both parts of Questions 2, 3 and 4 (36%, 32% and 18% respectively). Furthermore, the medians for the seven judgments in Questions 1 to 4 differed markedly, as did the size and direction of differences between the pairs of judgments.

The next question was whether an amalgamation of the respondents' conditional likelihood estimates, along Bayesian lines, would yield a decision that related logically to the verdicts that they then gave. In contrast with the *Adams* case, in this study the respondents were *not* shown how to derive likelihood ratios or how to amalgamate the weight of the non-DNA evidence. This was done by the researchers and not communicated to the respondents. In all cases, processing the estimates in this way significantly reduced the extreme probability of one in 2 million posited by the prosecution, bringing it more into line with the order of magnitude of the 'times'

judgments expressed in response to A(ii).

Only four respondents (2%) would have given odds in favour of innocence, so there is clearly no *direct* association with the verdicts chosen by the respondents (77, or 35%, still chose 'not guilty'). Certainly, the respondents did not appear to be arriving at the Bayesian result intuitively. This was further confirmed by the lack of 'logic' exercised by the 84 respondents who answered A(ii), the 'times' question. Nor did it seem that the respondents, as a group, were being particularly influenced by any one piece of non-DNA evidence. The individual likelihood and likelihood ratio estimates were generally poor predictors of the chosen verdict. Only their judgments about whether the rapist was likely to be a local man were significantly (0.05 level) associated with the final verdict.

In fact, few respondents actually changed their earlier verdict of 'guilty' or 'not guilty', despite being led through the process of quantifying the non-DNA evidence. By far the strongest single predictor of the final verdict was the initial verdict (significant at the 0.00001 level and successfully predicting 85% of the final verdicts). Although still capable of being a significant predictor, the *absolute* comparison between belief in likelihood of guilt and stated burden of proof only succeeded in predicting 67% of the final verdicts. 21.5% of the respondents were illogical in their initial verdict, if using this criterion. 33.3% were illogical in their final verdict. Most of this increase is accounted for by those respondents who should logically have found the accused 'not guilty'. 44% of this group of 120 respondents were illogical in their choice of final verdict, but more particularly 14% changed from an initial verdict of 'not guilty' to one of 'guilty'. We cannot know why they made this change, but possibly something in their rethinking and quantifying of the non-DNA evidence may have influenced them to do so.

## DISCUSSION

Making a comparison of likelihoods, given DNA evidence like one chance in 2 million of the accused being innocent, seems to be particularly difficult for lawyers, and it is likely that jurors will have similar problems. Firstly, it requires them to handle very large numbers. Koehler, 1997, also discusses this, referring for example to the need for jurors to comprehend 1 in 57,000,000,000 in the O J Simpson case. Secondly, they must appreciate that, even though it has no numbers associated with it, the non-DNA evidence *might* be capable of offsetting the massive-sounding number presented by the prosecution. There was certainly evidence from the present study that the respondents may have had

very general and basic difficulties in handling ratios and percentages. Even the fundamental appreciation by lawyers and jurors that one chance in 1 million is greater than one chance in 2 million is not something that can be taken for granted. However, there did seem to be evidence that most of the lawyers were able intuitively to compare two ‘quantities’, whatever their form (numerical or verbal descriptors) and make a logical decision about guilt or innocence, provided that the comparison was formulated in *absolute* (greater or smaller than) rather than in *relative* (times greater or smaller than) terms.

As Eggleston (1983) said, *some* mistakes are inevitable in dealing with uncertainty. In view of the importance of ensuring *optimal* legal decision-making therefore, the authors are inclined to concur with Koehler (1992) that more research is needed into the decision-making process, and also into how to educate lawyers and members of the general public so that they will have greater facility with matters involving uncertainty. It seems that, even though the respondents appeared to be able to make likelihood estimates related to the non-DNA evidence, these did not appear to impact significantly on their choice of verdict. The following research question remains open. If the respondents had been shown how to amalgamate their likelihood estimates, as were the jurors in the *Adams* case, would this have made a difference to their verdicts, by giving them an explicit numerical answer to set against the prosecution evidence?

If Donnelly is correct in saying that this is ‘the *only* logically consistent way’ of balancing the DNA and non-DNA evidence, then lawyers and jurors must be taught how to do this. On the other hand, perhaps the Appeal Court judges in the *Adams* case have a point when they suggest that the jury’s ‘individual common sense and knowledge of the world’ is what prevails. The evidence from the present study would indicate that this kind of absolute (greater or smaller than) comparison did produce more *coherence* with their verdicts. Relative (ratio, times greater or smaller than) comparisons did not seem to be so well handled by the respondents. Of course, coherence is not necessarily synonymous with a *correct* verdict, and the frequency with which UK verdicts are being declared ‘unsafe’ would suggest that their Lordships should be less sanguine about relying on the jurors producing the right outcome by applying ‘absolute’ common sense.

## REFERENCES

Eggleston, R (1983, 2nd edition). Evidence, Proof and Probability. *Weidenfeld and Nicolson*. ISBN 0 297 78263 0.

Hawkins, P and Hawkins, A (1998, in press). Probability Misconceptions of Lawyers and their Implications for Legal Education. *Legal Studies*.

Koehler, J (1992). Probabilities in the courtroom: and policies. In Kagehiro, D K and Laufer, W S (Eds). *Handbook of Psychology and Law*. Springer-Verlag. Chapter 9, 167-184

Koehler, J (1997). One in Millions, Billions and Trillions: Lessons from People v Collins (1968) for People v Simpson (1995) *Journal of Legal Education*, 47(2), 214-23.

R v Adams [1996] 2 Cr. App. R. 467.

R v Adams (1997). The Times 3 November 1977.