Forthcoming ***Journal of Law, Economics and Organization***

# Errors in Judicial Decisions: Experimental Results

Joep Sonnemans*
Universiteit van Amsterdam

Frans van Dijk**
Council for the Judiciary

**August 2010**

**Abstract**
In criminal cases the task of the judge is foremost to transform the uncertainty about the facts into the certainty of the verdict. An extensive literature shows that people deviate from rationality when dealing with probability. It seems therefore unavoidable that in difficult criminal cases miscarriages of justice occur, but this is hard to study in the field. In a laboratory experiment we examine the relationship between evidence of which the diagnostic value is known, subjective probability of guilt and errors in verdicts for abstract criminal cases. We look at two situations: (1) all evidence is given and (2) evidence can be acquired. In both situations verdicts are inaccurate. For given evidence, errors are biased towards the most serious type, unfounded conviction. In the situation where evidence can be acquired, participants do not acquire enough which results in many mistakes, evenly divided over unfounded convictions and unfounded acquittals. We suggest ways to reduce error.

JEL codes: C91, K14, K40

---

* CREED, Amsterdam School of Economics, Universiteit van Amsterdam, Roetersstraat 11, 1018WB Amsterdam, The Netherlands and the Tinbergen Institute. Email: J.H.Sonnemans@uva.nl
** Council for the Judiciary, Kneuterdijk 1, 2509 LP The Hague, The Netherlands. Email: F.van.Dijk@rechtspraak.nl
We thank two referees and the editor for very helpful remarks, the Study Centre of the Netherlands Judiciary for securing the cooperation of candidate judges and the priority research area Behavioral Economics of the University of Amsterdam and the Council for the Judiciary for financial support.

## 1. Introduction

Fact finding is central to the adjudication of criminal cases. Core business of judges is to transform uncertainty about the facts into the certainty of the verdict. While questions of law may arise, the dominant issue is generally what happened and who did it. This implies the evaluation of evidence, in particular with respect to the likelihood of plausible courses of events. To reach a verdict all pieces of evidence have to be combined to conclude whether or not the total burden of evidence meets a relevant criterion, such as in the US 'beyond reasonable doubt' or in the Netherlands 'lawfully and convincingly proven'. In most cases all or nearly all evidence points in the same direction, which makes the verdicts relatively easy. However, when evidence is weak and/or contradictory decision making is hard, and errors are likely to occur. In many of these hard cases important evidence is delivered by forensic experts and by its nature that evidence is probabilistic. However qualitatively, the judge has to combine these probabilities to estimate the likelihood of guilt.

From a decision-theoretic perspective (see e.g. Pratt et al., 1995), judicial decision making is a case of decision making under uncertainty, and can be modeled as such (e.g. Schrag and Scotchmer, 1994 and Lando, 2006). Optimal decisions require the correct application of the rules of probability. These rules are well understood, but hard to apply in practice and often counter intuitive (see Dawd, 2005). Thinking in probabilities does not come easy to humans, and it has been suggested the more so for law students due to selection and training (Wagenaar, 2006). Two major sources of imprecision can be distinguished: non-systematic deviations from optimality due to among others performance errors and computational limitations, and systematic divergences, as documented by the extensive literature on biases. We refer to the review by Stanovich and West (2000), which concludes that in general both sources are important. Guthrie et al. (2001 and 2007) show that, like other humans, judges are subject to biases.

It seems unavoidable that the aforementioned hard decisions will often result in mistaken verdicts, in the sense of either unfounded convictions or unfounded acquittals. Not much is known about this. The literature on biases and other deviations from rationality, which we will discuss in section 2, focuses on the occurrence of the deviations, and not on their impact on (judicial) decisions. To some extent this is understandable, as judicial practice

does not lend itself easily for scientific investigation. Whether verdicts are correct or mistaken cannot in general be established independently. Jurisprudence sometimes shows faulty reasoning, such as entirely discarding weak evidence during the process of combining all evidence, but that does not necessarily imply that verdicts are wrong. Only in exceptional cases it becomes clear that a miscarriage of justice has occurred, for instance because the real perpetrator turns up or new forensic methods make it possible to re-examine cases, but that in itself does not prove that errors were made. Field research into judicial error is therefore inherently problematic and recourse must be taken to experimental methods. Another issue is that, while judicial practice is all about uncertainty, it generally does not confront uncertainty explicitly. Probabilities are not assessed quantitatively, and the rules of probability are not applied, at least not in a transparent manner. The judge must be convinced of the guilt of the defendant, but what it takes to be convinced is not made explicit. Judicial decision making is generally qualitative and often intuitive (see Guthrie et al, 2007), and thus far apart from the theory of decision making under uncertainty with its quantitative and rational orientation, and its game theoretic extensions[1]. These approaches seem to belong to different worlds. The fundamental issue at stake here is whether the current practice of judicial decision making measures up against the normative model of decision theory, but also how to establish this.

The traditional qualitative instead of quantitative approach of judicial decision making can be explained by the lack of precise quantitative evidence in the past. However, this is rapidly changing. The trend in forensic evidence is that insight in the uncertainty involved is improving. In the eighties it was suggested that mathematical evidence was likely to increase dramatically in the near future (Jonakait, 1983). This has proven to be true, and the trend goes on unabated. Knowledge about the magnitude of uncertainty does not only pertain to technical evidence such as DNA-analysis, but, increasingly, also to classical evidence such as fingerprint and handwriting with improving digital pattern recognition (e.g., Brink et al., 2007). Also, insight into the diagnostic value of non-technical evidence like multiple recognition and confessions is improving (e.g., Clark and Wells, 2007, Kassin et al., 2007). Consequently, judges can dispose of more precise information, which lends itself or even requires to be combined quantitatively by using the rules of probability. This development makes it easier to apply the normative model. The research questions involved in addressing this issue are manifold, and encompass among others legal, cultural and cognitive aspects. We will not

address legal questions, such as whether the laws of evidence and procedure stand in the way of applying the normative model. Also, we will not look at judicial culture such as the habits and customs among judges in handling cases. We will focus on cognition, and by experimental methods examine the question with what precision highly educated people, including candidate judges, are actually able to handle probabilistic decision problems that share the uncertainty characteristics of real criminal cases judges are confronted with, and, consequently, to what extent errors of different types are made in comparison with the normative model. In order to enable a sharp comparison between actual behaviour and normative model, the decision problems are simplified in the sense that all necessary quantitative information is available to solve the problems uniquely, when applying the normative model. This simplification entails that we focus exclusively on the interpretation and combination of evidence of which the diagnostic value (further denoted as 'strength') is completely known. We will try to answer the question to what extent people measure up to the normative standard of fully rational and cognitively unconstrained optimization under risk neutrality for different combinations of evidence. These combinations vary from strong and consistent evidence to weak and contradictory evidence. We are also interested in whether it helps to have a legal, science or social science background. From a methodological perspective we would like to point out that in an integrated manner we examine coherence (alignment with rationality) and correspondence (actual accuracy) of decisions, as for example discussed by MacCoun (1999).

At the onset we want to point out that in order to study decisions the incentives of subjects need to be predetermined. This implies that the cost of errors needs to be specified and actually incurred by subjects. Using standard methods of experimental economics, we do this in monetary terms. This makes it also possible to evaluate the cost of error in monetary terms.

The experiment consists of two parts. In the first, all evidence is externally given and participants have to decide cases immediately (part 1: verdict) and also report their subjective beliefs that the defendants are guilty. In the second part only a single piece of evidence is given and participants can acquire further evidence, before they reach a

verdict (part 2: inquiry and verdict). Part 1 is the stepping stone to part 2, but may also be closer than part 2 to some legal traditions such as that of the US, in which judges are largely passive in the sense that only parties can provide evidence (e.g., Way, 2003). Part 2 is closer to most European traditions, in which the judge actively leads the inquiry in court. As experimental economics shows, search compounds the difficulties already noted, as people have a tendency not to search optimally (e.g. Sonnemans, 1998, and the references therein). We draw the results together to suggest ways to reduce error, especially those of the worst kind, unwarranted convictions.

## 2. Overview of literature

Evaluation of evidence requires thinking in terms of conditional probabilities: what is the probability that a defendant committed the crime, given the evidence and all other relevant information? Using Bayes' formula (see e.g. Mood et al., 1974), the information contained in the evidence can be combined with the initial belief of the judge about the guilt of the defendant to arrive at a new assessment of his guilt. In terms of prior and posterior odds, where *g* stands for guilty, *ng* for not guilty and *e* for evidence and with $P(g|e) + P(ng|e) = 1$:

$$\frac{P(g\,|\,e)}{P(ng\,|\,e)} = \frac{P(g)}{P(ng)} \bullet \frac{P(e\,|\,g)}{P(e\,|\,ng)} \qquad (1)$$

In words: Posterior odds = Prior odds * Strength of evidence

If the ratio $P(e|g)/P(e|ng)$ is larger than one, the evidence is incriminating. If the ratio is smaller than one, the evidence is exonerating. $P(g)/P(ng)$ is the initial belief (prior odds) and $P(g|e)/P(ng|e)$ the adjusted belief, given the evidence (posterior odds). Using the identity $P(g|e)$ can be expressed in the terms of the right hand side.

The literature shows that people have difficulty thinking applying these concepts. They have problems interpreting the individual probabilities of the right hand side of equation (1), and they have trouble combining these probabilities correctly. In several experiments it has been found that compared to this rule participants do not give sufficient weight to the evidence. In one form or the other, participants were given or first asked to assess $P(g)/P(ng)$, then given further evidence of known strength and asked to assess $P(g|e)$. For instance, Thompson and Schumann (1987) let participants assess the

probability of guilt on the basis of a case description, and then gave them a further incriminating piece of evidence. They found that the Bayesian posterior probability was significantly higher than the participants' subjective assessment of guilt. And they concluded that “this finding is consistent with the general tendency of people to be more conservative than Bayes' theorem when revising judgments in light of new information”. They got the same result when they gave participants the prior assessment of guilt. Faigman and Baglioni (1988) and, in a different setting of liability for personal injury, Bornstein (2004) got this result as well. Other references are Edwards (1968) and Saks and Kid (1980). This literature suggests that participants tend to underestimate the strength of evidence.

In addition to this general tendency, people are subject to specific bias. There is a host of literature about representativeness bias, starting with Kahneman and Tversky (i.a., 1972). They have made plausible that people tend to use a simplifying heuristic to evaluate probabilities. Applied to our context, the representativeness heuristic refers to a tendency in decision makers when they assess the probability that a defendant is guilty to base their judgments on the extent to which the evidence available is representative of guilty behavior. Guthrie et al. (2001) give the demeanor of the defendant as example. If the defendant is nervous and shifty, it will be seen as evidence of guilt. When he appears at ease, this will be seen as evidence of innocence. This leads astray if the prevalence of nervous and shifty behavior among innocent defendants or of at ease behavior among guilty defendants is not considered. In extreme form this bias leads to the so called inverse fallacy: the probability of guilt given the evidence is equated to the probability of evidence given guilt: $P(g|e) = P(e|g)$. Note that this also implies that initial beliefs are ignored. The inverse fallacy is also documented in medicine (Eddy, 1982): the probability that a patient has a tumor, if a test result is positive, is equated to the probability that the test result is positive, given that the patient has a tumor. The results of such thinking can be disastrous. In the literature many examples of such bias are given. Here, we note only the influential experiment of Thompson and Schumann (1987), also because it points to the complexities of behavior. In catchy phrases, Thompson and Schumann call the overvaluing of evidence by not taking into account the a priori likelihood that a defendant is not guilty the Prosecutor's fallacy. However, they also find an opposite fallacy, which

they term the Defense Attorney's Fallacy and which signifies that probabilistic evidence is completely ignored. They find that sizable numbers of participants fall prey to one or the other of the fallacies, but that the Defense Attorney's occurs more frequently. When participants use the information, their assessment of guilt tends to be lower than Bayes' rule would stipulate, as we noted already. Thus, while some individuals may fall prey to the Prosecutor's fallacy, this is not the case for the whole group that does not disregard the evidence. Obviously, in Thompson and Schumann's experiment many deviations from the statistical rule occur. Representativeness bias is only one of them[2].

When it comes, more specifically, to the cognition of judges, Guthrie et al. (2001) have shown that they, like other professionals such as doctors, engineers and options traders, are prone to cognitive illusions. They examined five biases, among which the representativeness bias. The others were: anchoring (making estimates based on irrelevant starting points), framing effects (treating losses differently than equivalent gains), hindsight bias (perceiving past events to have been more predictable than they actually were) and egocentric biases (overestimating one's own abilities). They showed by means of a set of problems they asked a large number of US federal magistrate judges to solve that judges suffer from these biases. These judges performed better than other decision makers with respect to framing and representativeness. Focusing on representativeness, 41% was not subject to the this bias, while according to the authors in a comparable study about doctors only 20% gave correct answers. Nonetheless, 40% of the judges was way off, and gave answers consistent with the above mentioned inverse fallacy, implying that they overrate the evidence. It should be noted, however, that this study was not a controlled experiment. So far, we can conclude that judges are not above the biases that plague other humans.

Another idea that can be found in the literature is that exonerating evidence gets less weight than incriminating evidence. In an experiment in which they provided participants with single incriminating and exonerating evidence and combinations of such evidence, McAllister and Bregman (1986) found that “nonidentifications had less impact on perceptions of guilt than identification for both eyewitness testimony and fingerprint evidence.”(p168). Perception of guilt was measured by asking participants to rate the defendant's innocence or guilt on a nine point scale, and their confidence in the decision

on a similar scale. The authors explain this finding in terms of a general tendency for negative information to be given greater weight than positive information. Whatever the truth of this explanation, their finding links with the broader issue that in criminal investigations there is a tendency to report only incriminating and not exonerating results. See Clark and Wells (forthcoming) on eyewitness identification, which views the tendency to ignore the diagnostic value of nonidentifying witnesses as a form of a confirmation bias or tunnel vision. Their research demonstrates that all witness responses should be taken into account and not only that of the witness(es) who identified the suspect. This argument can be generalized to all criminal investigations: an inquiry that leads to nothing, unless irrelevant (see below), is informative about the possible guilt of a suspect.

All of the above focuses on systematic divergences from the normative response. Given the inherent difficulties in applying statistical concepts, it is likely that non-systematic deviations occur as well. Stanovich and West (2000) distinguish between performance errors, computational limitations, applying the wrong normative model and alternative task constructs. One can readily hypothesize that, while all these factors may play a role, computational limitations are particularly inevitable.

The above findings are interesting, but far from conclusive. The literature cited suggests different sources of error. The importance of these sources is, however, not consistently established. Also, the research does not address the issue whether the sources of error would actually lead to wrong decisions. While research may show wrong assessments of probability of guilt given the evidence, it still needs to be established that this leads to wrong verdicts. For instance, while people may underestimate the probability of guilt, they may convict at a lower probability threshold than rationality would require. The reverse may also be true. To address these issues a more integral approach of judicial decision making is needed which takes probability assessments and verdicts into account.

## 3. Conceptualization

An integral approach of judicial decision making must incorporate the following issues.

### Errors and incentives

From the perspective of the accuracy of judicial decisions, judges can make two types of error[3]:

1. Convict an innocent defendant, which of course is a grave injustice to the individual concerned, but also leaves the real perpetrator at large at the risk of repetition.
2. Acquit a guilty defendant, which is an injustice to victims or their surviving relatives and also leaves the real perpetrator at large at the risk of repetition.

Legal standards such as “beyond reasonable doubt” and “convincingly proven” provide guidance. These legal standards reflect the trade off between the two errors. However, the standards are unavoidably vague, and may be interpreted differently by judges and by the same judge over cases and over time.

Errors in the above sense do not necessarily imply mistakes for which judges are to be blamed. Occasionally, all evidence will inculpate an innocent suspect, and, more commonly, evidence against a guilty suspect may be insufficient to rule against him. Also, whether a person really committed a crime cannot be ascertained independently. As noted in the introduction, only in exceptional cases convicts are unequivocally exonerated, because the real perpetrator turns up[4] or the application of new technology makes a reassessment of the evidence possible. DNA-techniques are the obvious case in point. Nonetheless, judges will have a strong intrinsic motivation to avoid error, and also an extrinsic motivation. Their independence shields them from direct repercussions, but reputation can be affected negatively. After the high profile case in the Netherlands, mentioned in footnote 3, and the public uproar it caused, the conviction rate declined (van der Heide, van Tulder and Wiebrens, 2007**).** This suggests that judges became more aware of the repercussions of a miscarriage of justice, and, consequently, became more careful. Still, judges cannot spend unlimited time on cases, as other cases would be delayed and the criminal justice system would grind to a halt. Consequently, judges have an interest in concluding cases. This results in the following incentive structure (table 1).

| | | **Real situation the accused is** | |
|---|---|---|---|
| | | **the perpetrator** | **innocent** |
| **Verdict** | Conviction | $a>0$ | $b<0$ |
| | Acquittal | $c<0$ | $d>0$ |

**Table 1.** Benefits and costs of judicial decisions for the judge

From a legal perspective, *a* and *d* should be equal: the judge should be indifferent between these outcomes. It would seem likely that $b << c$. The weights judges attach to these outcomes are fundamentally implicit to their functioning and cannot be known with any precision. Therefore, we impose them in the experiment. Our results will depend on the comparison of actual with optimal decisions, and the numerical values as such are irrelevant.

In combination with the judge's attitude towards risk, the incentive structure determines the probability of guilt minimally needed to convict a defendant. The risk neutral optimal decision maker is indifferent between conviction and acquittal when $a\text{p}+b(1\text{-p})=c\text{p}+d(1\text{-p})$ with p the probability of guilt. Or: $\text{p} = (d-b)/(a-b-c+d)$.

**Evidence and uncertainty**

When a serious crime has been reported, investigations start. Depending on the legal system these investigations are supervised by a judge. We will not go into this process, and focus on a suspect being brought to court. The investigations will have led to sufficient evidence in the view of the prosecution to warrant the case to proceed to court.

| **Inquiry** | **Possible outcome** | **Probability of evidence if the accused is the perpetrator** | **Probability of evidence if the accused is not the perpetrator** | **Strength of evidence** |
|---|---|---|---|---|
| *i* | **Incriminating** | $x$ | $v$ | $x/v > 1$ |
| | **Exonerating** | $y$ | $w$ | $y/w < 1$ |

**Table 2**. Evidence resulting from criminal investigations, $x + y = 1$ and $v + w = 1$.

Table 2 explains how the strength of a piece of evidence is calculated. Note that irrelevant investigations will result in a neutral outcome in the sense that it makes no difference for the outcome whether a suspect is or is not guilty ($x/v = 1$). As convictions cannot be based on a single piece of evidence in most legal systems, the probabilities associated with different pieces of evidence generally have to be combined. The reality is that in some cases evidence will be contradictory. An example is the well documented case in the UK against Adams, in which DNA evidence conflicted with other evidence and in particular with a multiple recognition (see Donnelly, 2005).

Denoting the strength of a piece of evidence *i* as $E_i$, generalizing (1) gives for independent evidence:

$$Odds_{posterior} = Odds_{prior} * \prod_{i=1}^{n} E_i \qquad (2)$$

where: $E_i = P(e_i|g)/P(e_i|ng)$

The subjective assessment of the posterior odds by a judge may differ from the mathematically correct calculation, using the strengths he attaches to the individual pieces of evidence and his basic belief about the guilt of the defendant. Also, the strength the judge attaches to a particular piece of evidence may differ from an, as far as possible, objective assessment of this strength, for instance based on scientific research[5]. Initial beliefs (prior odds) may also vary. A judge may apply a presumption of innocence, may be influenced by his experience that most of the accused are guilty or may apply a more individuated criterion, dependent on his experience with specific crimes or perhaps his prejudices against certain suspects. Therefore, in the experiment the prior has to be specified and given to the participants.

It should be noted that by choosing a basic hypothesis structure (defendant is guilty or not guilty) and independence of the evidence we focus on the bare essentials of judicial decision making. Hypothesis structures can of course be much more complex and evidence can be interdependent (see e.g. Schum and Martin1982, Martin and Schum 1987 and Robinson and Hastie 1985).

**Decision and search**

When only evidence presented by the prosecution and the defense is allowable, the task that remains for the judge is to decide the case[6]. This requires him, however qualitatively and intuitively, to evaluate his subjective assessment of guilt against a threshold for conviction, as discussed above. The risk neutral optimal decision maker applies equation (2) to calculate the probability of guilt, given all information available to him, and compares that with the threshold probability as calculated above.

Externally given evidence does not capture the complexity of judicial decision making, when judges play an active role in hearing cases, as happens in inquisitorial legal systems and in some adversial systems as well (see Way, 2003). The prosecution and the defense present evidence, but the judge questions (expert) witnesses, decides to hear other witnesses or that further investigations need to take place. In this context the judge has to decide when to stop the investigation in court. He then has to rule. This brings in a further complication, because the judge has to weigh the probable reduction of uncertainty by further inquiry into the case against the time and effort this requires of him and other parties and the resulting delay of cases on the docket. Again, this is a highly subjective decision. In part 2 of the experiment this decision is controlled by allowing participants to acquire pieces of evidence at fixed costs. It should be noted that the threshold probability calculated above does not apply anymore. The amount of evidence a risk neutral optimal decision maker acquires now depends on the consistency of the evidence. When all evidence points towards guilt or all evidence points towards innocence, the decision maker will stop sooner acquiring further pieces of evidence than when the evidence is contradictory. A general solution to this decision problem cannot be given. In 5 we specify the evidence and give the corresponding solution.

Note that we have simplified the decision problem also in another way. We abstract from the roles prosecutor and defense play and their incentives, for instance, to selectively present evidence (see Landes and Posner, 2001 and footnote 1). Our experiment can be seen as a benchmark for future extensions that allow for strategic interaction among the parties involved.

## 4. Research questions

Section 3 provides the general framework to compare experimentally the decisions of actual decision makers with risk neutral optimal decisions, and to analyze the causes of differences. We can now formulate the research questions to be answered by the experiment more specifically. The first four questions deal with the situation of given evidence. *First*, to what extent are decision makers able to reach accurate verdicts, given evidence and incentives? By accurate we mean that verdicts are close to the outcome of the normative model. *Second*, do decision makers decide the cases in the manner the normative model prescribes, i.e. form beliefs about the probability of guilt and on that basis reach verdict, or do they proceed in a more intuitive manner? *Third*, in as far as decision makers form beliefs about probability, to what extent does the combined, subjective probability individual decision makers attach to the total burden of proof differ from the objective probability, given the strength of each piece of evidence? In view of the literature discussed we would expect subjective probability to be lower than objective probability in case of stronger incriminating than exonerating evidence. *Fourth*, again in as far as decision makers form beliefs about probability, to what extent differ the verdicts from the normative model, given their beliefs about the probability of guilt? We can then conclude whether wrong decisions are foremost caused by the subjective assessment of aggregate probability or, given aggregate subjective probability, by the rulings. *Fifth*, when participants can acquire evidence, to what extent differ their verdicts from optimal decisions? Are the differences foremost caused by not acquiring the optimal amount of evidence or by wrong verdicts, given the collected evidence? From the economic literature we expect that many participants will not search long enough. *Sixth*, does it matter whether participants have a background in law, science or social sciences for both situations with respect to the evidence?

## 5. Design

Computer screens and the instructions are available in the downloadable appendix and the reader can anonymously participate in an online version of the experiment at www.creedexperiment.nl/recht2/begin.html.

All participants participated in two experiments. In the first small experiment their attitudes towards risk and loss were measured by means of lotteries (comparable with Holt and Laury, 2002). The main experiment dealt with judicial decision making and consisted of the two parts already explained. In part 1, denoted 'verdict', the evidence was given, and participants just had to decide the cases. In part 2, denoted 'inquiry and verdict', evidence could be acquired by ordering inquiries. In both parts participants had to decide 30 cases, with which they could earn money. In each part it was possible to make losses. In addition to the earnings to be discussed below, all participants earned a salary of 100 points (equaling 1 euro) per case in both parts. Eventual losses in each part were subtracted from the salary in that part with a minimum earning of 0 per part[7].

Participants were informed in advance that in about 15 of the 30 cases the defendant was guilty, so the a priori odds were 1. The 30 cases of both parts are given in the Appendix.

To guarantee their understanding of the experiment, participants had to answer computerized questions and received feedback. A participant could only continue if (s)he had answered the questions correctly. Then the participant had to continue with 6 practice cases, with which no money could be earned. Feedback was given per practice case and after all the practice cases, and included the pay-off if the case(s) had been for real. The outcomes of the 30 cases of both parts were given at the end of the experiment.

**Part 1: verdict**

We used the following structure of the evidence, which was given and explained to the participants. Three types of investigations are distinguished, each resulting in either incriminating or exonerating evidence. In a case, several inquiries could take place, also of the same type.

The procedure to generate the 30 cases and associated evidence was as follows. First, whether the defendant was guilty or not was randomly determined within the constraints mentioned above. Second, it was randomly determined which investigations would take place (type 1 and 2 with 30% probability, type 3 with 40%). Third, the outcome of each investigation was determined randomly from the probability distribution, dependent on the guilt or innocence of the defendant, as given by table 3. In

this way 3 to 6 pieces (all equally likely) of evidence were generated. The evidence was presented sorted by kind (incriminating or exonerating) and strength.

| Type of inquiry | Possible outcome | Code in experiment | Probability of evidence if the accused is the perpetrator | Probability of evidence if the accused is not the perpetrator | Strength of evidence |
|---|---|---|---|---|---|
| 1 | **Incriminating** | **1INC** | 84% | 36% | 84/36=7/3=2.33 |
| | **Exonerating** | **1EXO** | 16% | 64% | 16/64=1/4=0.25 |
| 2 | **Incriminating** | **2INC** | 64% | 16% | 64/16=4.00 |
| | **Exonerating** | **2EXO** | 36% | 84% | 36/84=3/7=0.43 |
| 3 | **Incriminating** | **3INC** | 60% | 40% | 60/40=3/2=1.50 |
| | **Exonerating** | **3EXO** | 40% | 60% | 40/60=2/3=0.66 |

**Table 3**: Strength of evidence as used in part 1, verdict.

For every case, participants reported the subjective probability that the accused was guilty, and made the decision to convict or acquit. Either the decision or the subjective probability was rewarded (both with equal probability): the decision according to table 1 with, in points, $a$=$d$=100, $b$=-1500 and $c$=-300, and the belief according to a quadratic scoring rule. The reader is referred to Appendix 1. The scoring rule is incentive compatible for risk neutral individuals (see Offerman et al., 2008). This procedure prevents hedging behavior by participants[8]. All participants received the same cases and evidence.

With the chosen parameters the risk neutral optimal decision maker is indifferent between conviction and acquittal when the probability of guilt, given all evidence, is 0.8 (see section 3). Thus, this decision maker should only convict the accused when the evidence points to a probability of guilt higher than 80%. This occured in 8 cases. Note that the parameters are set in such a way that participants have a very strong incentive not to convict innocent defendants. Still, this probability is lower than in practice would be the case. For example, in an experiment Martin and Schum (1987) asked subjects to assess the threshold for “beyond reasonable doubt” and found 91% for most crimes and

99% for murder. We gave more weight to reliable data collection than superficial realism in this respect.

**Part 2: inquiry and verdict**

Only one piece of evidence was given. Participants had the option to order inquiries. All inquiries were of the same type (type 2 of table 3). Each inquiry either resulted in an incriminating or exonerating piece of evidence. In total six inquiries could be ordered. Acquiring a piece of evidence cost 10 points. Because this experimental situation is more complicated, participants were not asked to report their subjective probability of guilt, but only to decide the cases.

In each case, a participant had to decide first whether or not to order an inquiry. If not, he had to decide the case with the verdict guilty or not guilty. If he decided to order an inquiry, he, subsequently, had to decide on a further inquiry, and so on. Figure 1 gives the optimal strategy. All participants received the same cases; the optimal decision maker would convict 12 of the 30 defendants, of whom 2 would be innocent. Note that in part 1 it is optimal to convict a defendant on the basis of two incriminating pieces of evidence of type 2 and no other evidence, while in the search part it is optimal to bear the small cost of acquiring additional evidence and reduce uncertainty further.

**Participants**

In order to test the influence of background, participants were enlisted from four groups: candidate judges and law, science and social sciences students. In the Netherlands, candidate judges are lawyers who passed a selection process and for six years occupy different functions in courts before becoming full judges. During these six years they have to follow courses and pass examinations. We ran the experiment during one of these courses. Most law students who participated were so called "honors' students". These students are the top 10% of their year. We do not consider this a validity threat, because most judges are recruited from the top segment. The social science group consisted of economics and psychology students. To facilitate participation the experiment took place at the regular Creed laboratory, at a computer room of the law school and at the study

centre of the Netherlands judiciary. In total 54 candidate judges and 162 students participated in the experiment (see table 6 and 7).

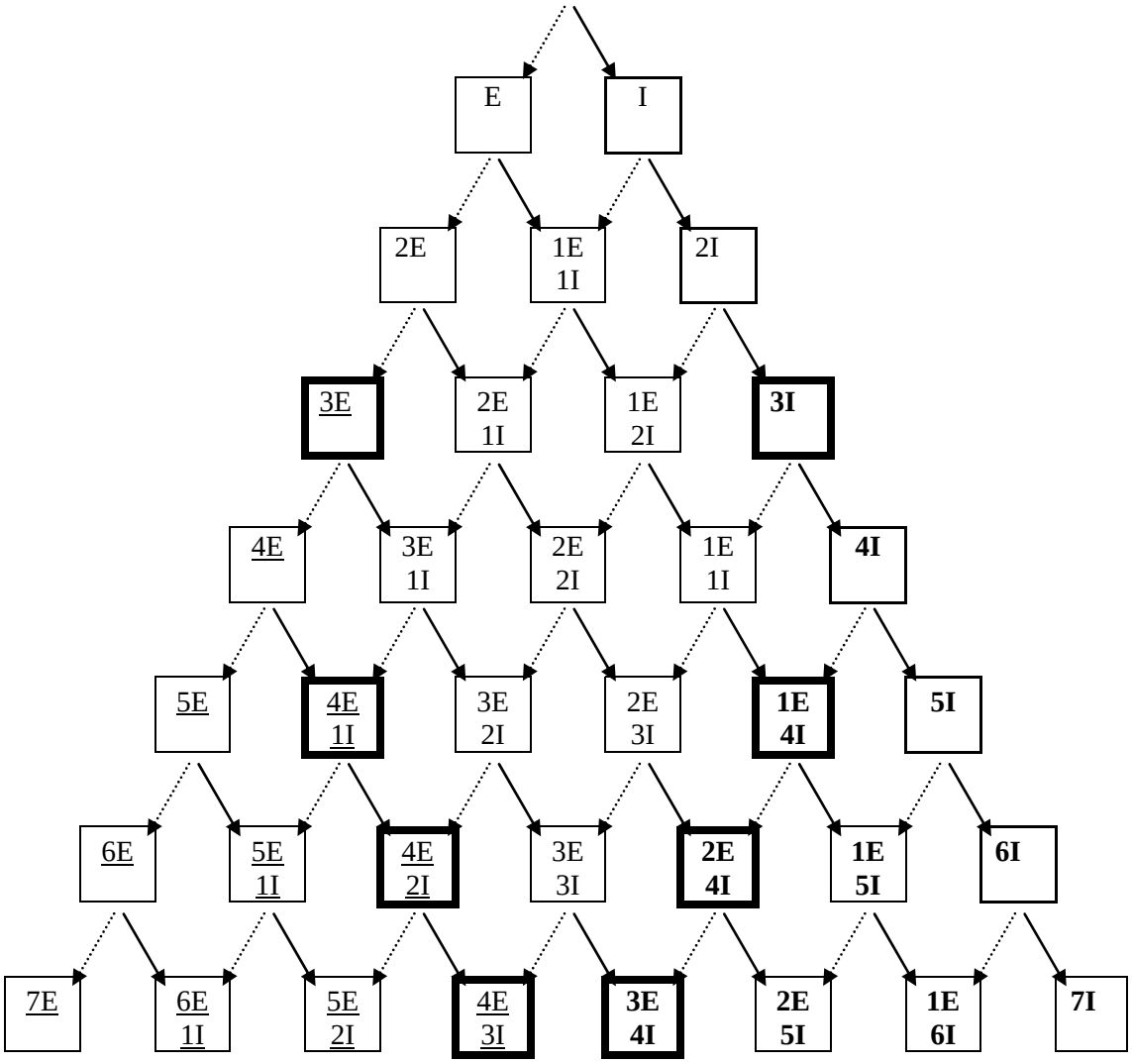


**Figure 1**. Outcomes and optimal strategies in the search part.

Note: the first piece of evidence is provided for free and the participant can buy sequentially up to 6 extra pieces of evidence. An incriminating piece of evidence is coded as I and a continuous arrow; an exonerating piece of evidence with E and a broken arrow. Combinations of evidence for which the optimal decision is to convict/acquit are printed bold/underlined; in other cases it is optimal to acquire more evidence. The situations with a bold border are the only optimal decisions (e.g., the best decision in 4E is to acquit the suspect, but this is not optimal because one should have stopped earlier and acquitted at 3E).

## 6. Results

### 6.1 Results part 1: verdict

The analyses of part 1 will be presented in four subsections. In the first we look at the overall relationship between given evidence and verdict. In the second, we examine whether participants' behavior is consistent with the normative model, which starts by assessing the probability of guilt and on that basis reaches the verdict. Consistency requires participants at least to form reasonably correct beliefs about the probability of guilt, given the evidence. In the third part, the relationship between belief and verdict is examined for those to whom the normative model applies. Finally, the errors will be analyzed. In that subsection we will also look at the impact of differences in background of participants (candidate judges, law, science, economics and other social sciences students).

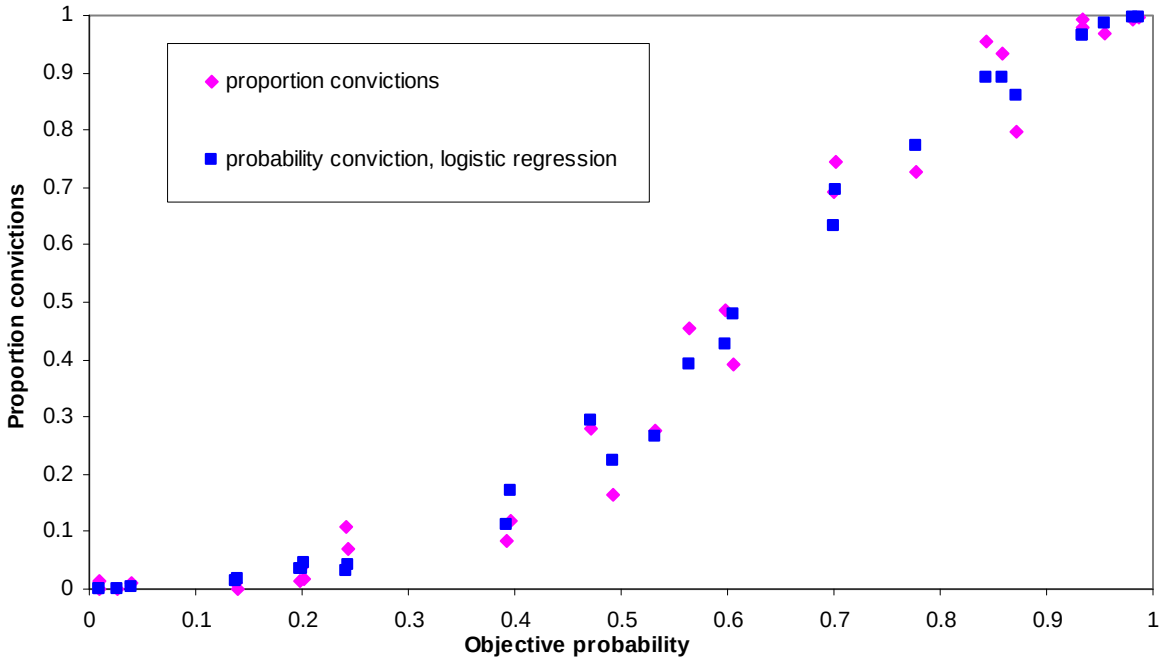


**Figure 2**. Proportion of conviction (diamonds) and the predicted proportion according to a logistic regression (squares) for the 30 cases.

6.1.1 Evidence and verdict

Figure 2 shows a scattergram for the thirty cases with on the horizontal axis the objective probability of guilt and on the vertical axis the proportion of convictions. Note that if all participants would be perfect-Baysian value maximizers, convictions would be observed if and only if the objective probability is higher than 80%. Actually, we do not observe this large step at 80%, but a gradual increase of convictions when objective probability increases.

We estimated a logistic regression with the decision as dependent and the frequencies of the different kinds of evidence as independent variables. The probability of conviction is estimated as $1/(1+e^{-Z})$ with

$Z=-1.04+1.56*INC_1+2.17*INC_2+0.67*INC_3-2.34*EXO_1-1.53*EXO_2-0.62*EXO_3$

in which the variables $INC_1$, $INC_2$, etc, stand for the number of times evidence of type 1INC, 2INC, etc, has been found in a particular case. All parameters are statistically significant ($p<0.0001$). The model predicts 87.5% of the decisions correctly. The estimates of the model are also presented in figure 2. We find no bias in the direction of relative underweighting of exonerating or incriminating evidence: the regression parameters of $INC_1$, $INC_2$ and $INC_3$ are approximately equal to the parameters of respectively $EXO_2$, $EXO_1$ and $EXO_3$ (with of course a change of sign). This means that incriminating and exonerating evidence of the same strength cancel each other. The constant in the regression of -1.04 implies that a priori (or when incriminating and exonerating evidence have the same strength) the participants will convict in about 25% ( $1/(1+e^{1.04})$ ) of the cases.

6.1.2 Evidence and subjective probability

It is likely that participants have different behavioral strategies or make errors of different size. The analysis in the previous subsection had to neglect this because the 30 binary decisions are too few to do a separate logistic analysis for each participant. However, we also asked for the probability of guilt, and this continuous variable can be analyzed on an individual basis (although the statistical power will be low with only 30 observations). To study how subjective probability of guilt relates to the available evidence, we calculate for each participant a log-linear regression with the reported subjective odds as dependent

and the number of pieces of evidence of each type as independent variables. The background of the analyses is as follows.

First, we transform the reported probability to odds. If the participant is a Bayesian updater, the resulting subjective posterior odds are the product of the prior odds (initial belief of guilt) and the combined strength of the evidence, as given by equation (2):

$$Odds_{Subj} = Odds_{\mathrm{Prior}} * 2.33^{INC_1} * 4^{INC_2} * 1.5^{INC_3} * 0.25^{EXO_1} * 0.43^{EXO_2} * 0.67^{EXO_3} \qquad (3)$$

with $INC_1$ the number of evidence provided of type 1INC, $INC_2$ the number of evidence provided of type 2INC, etc. If we take the logarithm of the odds, the formula becomes linear:

$$Ln(Odds_{Subj}) = \ln(Odds_{prior}) + INC_1 * \ln(2.33) + INC_2 * \ln(4) + INC_3 * \ln(1.5) + EXO_1 * \ln(0.25) + EXO_2 * \ln(0.43) + EXO_3 * \ln(0.67) \qquad (4)$$

Because the odds of the prior equal 1 (guilty and not guilty are equally likely), the $\ln(Odds_{prior})$ is 0. In other words, if we run a log-linear regression, we should find a constant of 0 and estimated parameters equal to the logs of the strength of each type of evidence.

The results of all regressions are displayed in the (downloadable) appendix. For 15 participants no regression could be calculated because of too little variation in the reported probabilities (for example, reporting 50% for all cases). In addition 19 participants clearly misunderstood the task and reported their confidence in their verdict instead of the subjective probability of guilt. In these cases all coefficients of the EXO variables are positive instead of negative. We divide the remaining 182 participants in two categories. Category 1 consists of the (118) participants for whom the regression works quite well. We use as (admittedly subjective) criterion that the adjusted R-square is at least 0.50 and not more than one coefficient has the wrong sign. Category 2 consists of the (64) participants for whom the regression makes less sense (for example, they report high probabilities in some - but not all - cases with mostly exonerating evidence where

they rightly acquit the suspect). Thus, only half of all participants consistently form expectations[9].

| | **Categories 1 and 2 (N=182)** | | | | | | |
|---|---|---|---|---|---|---|---|
| *Variable* | *B* | *SE B* | *Beta* | *T* | *Sig T* | *Subjective strength* $e^B$ | *Objective strength* |
| $INC_1$ | 0.67 | 0.03 | 0.33 | 26.36 | 0.00 | 1.95 | 2.33 |
| $INC_2$ | 1.04 | 0.04 | 0.31 | 25.72 | 0.00 | 2.84 | 4 |
| $INC_3$ | 0.27 | 0.02 | 0.13 | 11.31 | 0.00 | 1.31 | 1.5 |
| $EXO_1$ | -0.51 | 0.03 | -0.21 | -16.47 | 0.00 | 0.60 | 0.25 |
| $EXO_2$ | -0.38 | 0.03 | -0.17 | -14.66 | 0.00 | 0.68 | 0.43 |
| $EXO_3$ | -0.30 | 0.03 | -0.12 | -11.47 | 0.00 | 0.74 | 0.67 |
| (Constant) | -0.22 | 0.08 | -2.76 | 0.01 | | 0.80 | 1 |

| | **Category 1 only (N=118)** | | | | | | |
|---|---|---|---|---|---|---|---|
| *Variable* | *B* | *SE B* | *Beta* | *T* | *Sig T* | *Subjective strength* $e^B$ | *Objective strength* |
| $INC_1$ | 0.70 | 0.02 | 0.35 | 28.69 | 0.00 | 2.02 | 2.33 |
| $INC_2$ | 1.10 | 0.04 | 0.32 | 28.00 | 0.00 | 3.00 | 4 |
| $INC_3$ | 0.31 | 0.02 | 0.14 | 13.30 | 0.00 | 1.36 | 1.5 |
| $EXO_1$ | -0.72 | 0.03 | -0.29 | -24.29 | 0.00 | 0.49 | 0.25 |
| $EXO_2$ | -0.58 | 0.02 | -0.26 | -23.24 | 0.00 | 0.56 | 0.43 |
| $EXO_3$ | -0.36 | 0.03 | -0.15 | -14.18 | 0.00 | 0.70 | 0.67 |
| (Constant) | -0.16 | 0.08 | -2.08 | 0.04 | | 0.85 | 1 |

**Table 4**: Loglinear regression of the subjective odds with as dependent variables the frequencies of types of evidence. Top-panel: regression based on the data of 182 participants (categories 1 and 2, see main text). Adjusted R-square is 0.48
Lower panel: data of the 118 participants (category 1 only, see main text). Adjusted R-square is 0.68.

We also calculated a combined regression for the 182 participants in category 1 and 2 together (top panel of table 4) and for the 118 participants of category 1 (lower panel of table 4). For convenience the last two columns show $e^B$ and the objective odds. We find that the constant is slightly smaller (but not statistically significant) than 1,

meaning that a priori the participants consider the probability that the defendant is guilty a little less than 50%, actually about 44% (47% if only category 1 is considered). The coefficients for the different kinds of evidence are in the right order of magnitude, but too close to 1. This means that, in general, the strength of evidence is underestimated by the participants. This effect is smaller if we only consider the participants in category 1. The parameter for $EXO_1$ departs most from the norm in being given not enough weight.

Theoretically neutral combinations of exonerating and incriminating evidence (total strength of about 1) lead to subjective odds that are also close to 1, as long as 1EXO is not part of the evidence. First the exception: focusing on category 1, the two pieces of evidence 1EXO and 2INC should cancel each other exactly (the total strength is 0.25*4=1), but combine to a subjective strength of 3.00*0.49==1.47, incriminating instead of neutral. For medium strong evidence (1INC and 2EXO) the combination has subjective strength of 2.02*0.56=1.13, which is only slightly larger than 1: marginally incriminating. The objective neutral combination of weak evidence 3EXO and 3INC combines to a strength of 1.36*0.70=0.95, slightly exonerating. The combination "3INC 3INC 2EXO" leads to a subjective strength of 1.04 (objective strength is 0.96) and combination "3EXO 3EXO 1INC" leads to a subjective strength 0.99 (objective strength is 1.04).

Figure 3 displays the relation between objective and subjective probability; if all participants would be perfect Bayesians, all points would lie on the diagonal. In general the subjective probabilities tend to be less extreme; they are too close to 50%.[10] The pattern for category 1 looks very similar to the probability weighting functions as used in prospect theory (Tversky and Kahneman,1992).

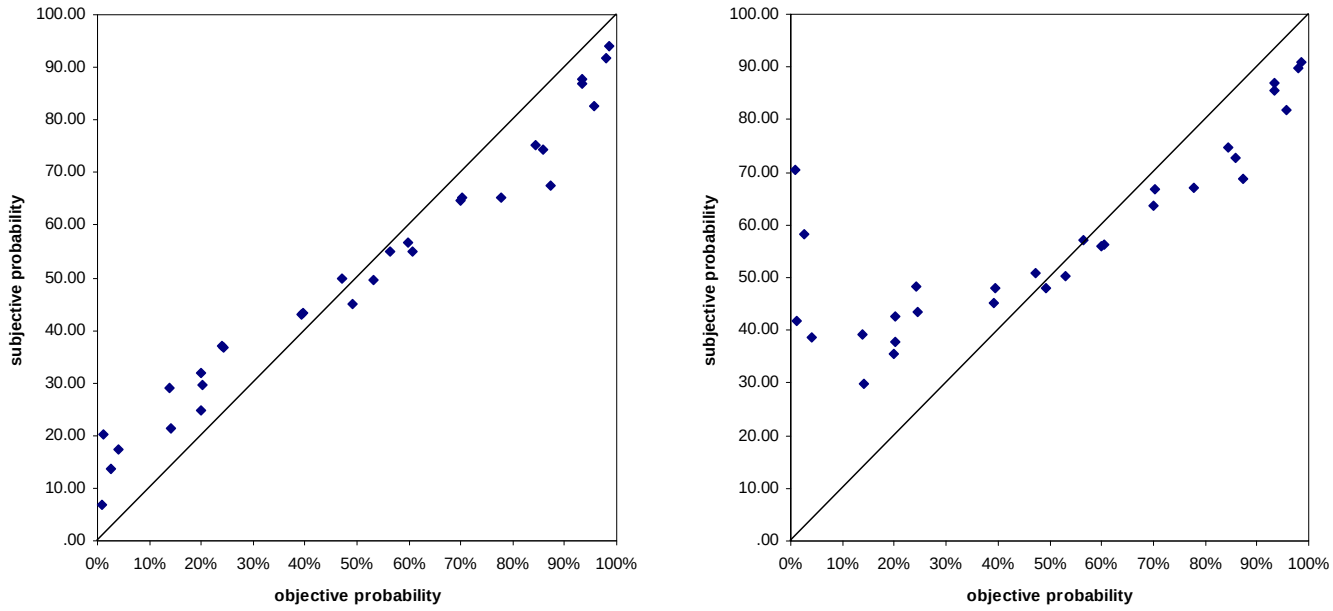


**Figure 3**: Scattergram of the average subjective probability for all 30 cases of part 1, with on the horizontal axis the objective probability. In the left panel the 118 participants of category 1, in the right panel the 64 participants of category 2.

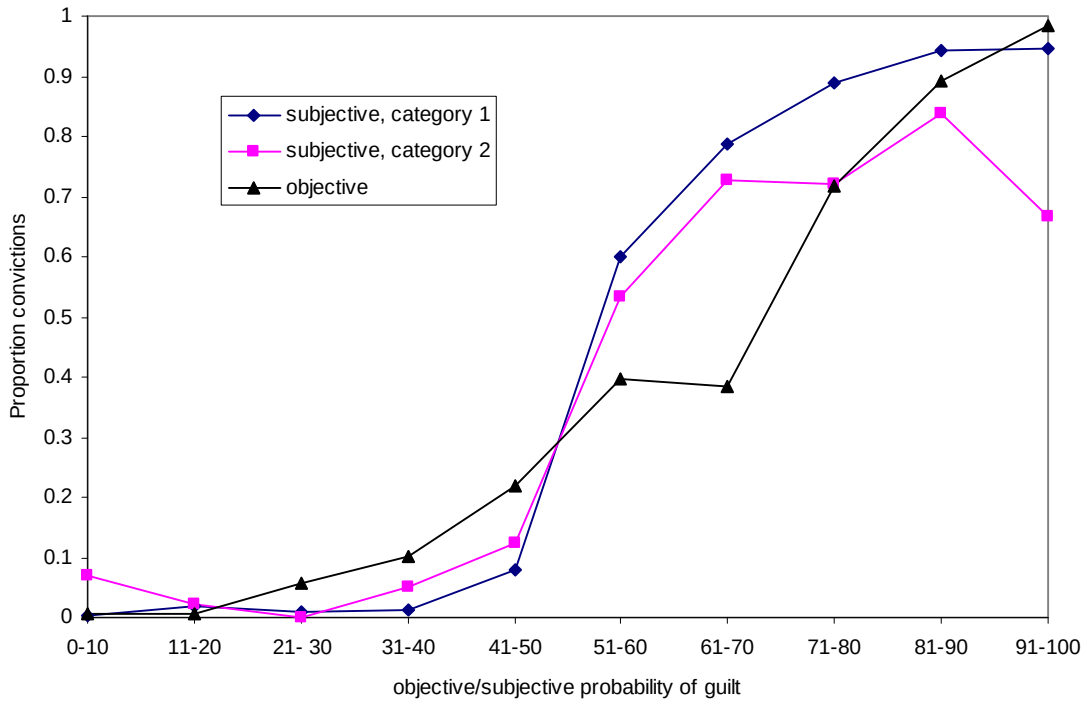


**Figure 4**: Average conviction rates by objective and subjective probability of guilt. For subjective probability a separate line is drawn for participants in category 1 and 2, for objective probability all participants are included.

| | | Errors | | | | | |
|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 7 | Total |
| Cutoff | 50 | 5 | 5 | 1 | 1 | 0 | 12 |
| | 51 | 1 | 2 | 0 | 0 | 0 | 3 |
| | 52 | 1 | 1 | 0 | 0 | 0 | 2 |
| | 53 | 0 | 0 | 0 | 1 | 0 | 1 |
| | 54 | 2 | 0 | 0 | 0 | 0 | 2 |
| | 55 | 12 | 3 | 1 | 0 | 0 | 16 |
| | 58 | 1 | 1 | 0 | 0 | 0 | 2 |
| | 60 | 15 | 10 | 4 | 0 | 1 | 30 |
| | 63 | 0 | 0 | 1 | 0 | 0 | 1 |
| | 65 | 3 | 5 | 2 | 2 | 0 | 12 |
| | 66 | 1 | 0 | 0 | 0 | 0 | 1 |
| | 67 | 0 | 1 | 0 | 0 | 0 | 1 |
| | 68 | 0 | 1 | 0 | 0 | 0 | 1 |
| | 69 | 1 | 0 | 0 | 0 | 0 | 1 |
| | 70 | 8 | 4 | 2 | 0 | 0 | 14 |
| | 72 | 0 | 2 | 0 | 0 | 0 | 2 |
| | 75 | 2 | 3 | 0 | 0 | 0 | 5 |
| | 78 | 1 | 0 | 0 | 0 | 0 | 1 |
| | 80 | 5 | 2 | 2 | 0 | 0 | 9 |
| | 100 | 0 | 1 | 0 | 1 | 0 | 2 |
| **Total** | | 58 | 41 | 13 | 5 | 1 | 118 |

**Table 5**: participants per cutoff point. A cutoff point is calculated for each participant such that the number of deviations is minimized. The table shows per cutoff point the number of participants and deviations. Category 1 only.

6.1.3 Subjective probability and verdict

Figure 4 shows the average conviction rate for categories of subjective probability. If all participants would be risk-neutral, they would only convict when the subjective probability is at least 80% and the graph would have a single step at 80%. Although the graph is increasing, we do not observe one single step.

Turning to individual behavior, for each participant the cut-off point that best fits the data is calculated. The results for the 118 participants of category 1 are displayed in table 5. About half of the participants (58) have followed consistently the same cutoff rule during all 30 decisions. It is plausible that the other participants adapted their strategy somewhat during the experiment; all except 1 deviate 3 times or less from their cutoff point, and can be considered largely consistent. The average cutoff point is 62.9%, lower than the risk-neutral optimum of 80% (if also category 2 participants are included

the same average cut off point is found, but with much more deviations)[11]. If subjective probability would be exactly the same as objective probability, this would mean that in general too many suspects are convicted. This effect is mitigated by the phenomenon that for the relevant cases (odds >1) participants on average underestimate the probability of guilt (figure 3).

Figure 4 also shows the average conviction rate by objective probability, and this line comes closer to the normative optimal decision rule[12].

6.1.4. Errors and individual differences

In 83.2% of the cases the decision equals the optimal decision. Of the 1089 deviations from optimality the majority is of the most serious kind: 1003 unfounded convictions. In what kind of cases are these important errors made? Figure 4 shows the average conviction rate by objective probability. The risk-neutral participant should only convict in the 8 cases where the objective probability of guilt is higher than 80%. In fact, in 95% of these cases suspects are convicted. Thus, the error of unfounded acquittal is very small. At the other extreme, in the cases where the evidence is very much in the direction of innocence (probability of guilt less than 40%) only 3.7% of the suspects are convicted. However, in the large area in between, where the evidence points in the direction of guilt, but it is not strong enough to rationally convict the defendant, much too many suspects are convicted. As a result, in the 9 cases that fall in this area the participants lost 9.20 euro on average (which most participants compensated by earning positive amounts in the other 21 cases). In 8 of these 9 cases, the average subjective probability is lower than the objective probability (although by only a few points). From this we can conclude that errors of unfounded convictions are not primarily caused by a wrong (too high) assessment of probability, but by a wrong decision based on reasonable beliefs.

Recent discussions about the difficulties that can arise if judges with a non-technical background have to make decisions based upon technical, probabilistic evidence is the motivation to compare participants with different backgrounds (table 6) The categories of participants that we constructed based upon reported beliefs about probability are not related to the background of the participants. We do not find statistically significant differences in the quality of the beliefs. As to the verdicts, subjects

with a background in law (candidate judges as well as law students) perform worse than others (2-sided Mann Whitney test $p<0.05$, $p<0.01$ and $p<0.01$ for comparison with science, economics and other social sciences respectively)[13]. Interestingly, the law and the science students use significantly more time per case (about 30 seconds) than the economics and other social sciences students (about 20 seconds)[14] and the candidate judges even more (36 seconds).

| **Subjects** | | **Assessment of probability** | | **Decision** | |
|---|---|---|---|---|---|
| | | **Average Error** | **N** | **Average Error** | **N** |
| **Candidate judges** | | 2.8 | 36 | 67.8 | 54 |
| **Students** | Law | 2.8 | 25 | 76.9 | 51 |
| | Science | 2.8 | 14 | 49.8 | 25 |
| | Economics | 2.8 | 28 | 53.2 | 54 |
| | Other social sciences | 3.1 | 15 | 49.5 | 32 |
| Total | | 2.8 | 118 | 59.4 | 216 |

**Table 6**: Average error per case, defined as the difference of expected earnings of actual decisions and expected earnings of optimal decisions, in cents, for participants with different backgrounds. For the assessment of probability only participants of category 1 are included.[15]

6.1.5 Discussion part 1 and further analysis

Apparently, probability concepts do not come naturally to most individuals (as many professors in statistics can testify). About half of the participants (118) report consistently subjective beliefs that are reasonable in the light of the available evidence. The other participants do not follow the theoretical path of evidence-belief-decision, but arrive at their decisions in different and, necessarily, more intuitive ways. We would expect that this behavior, which is farther removed from normative theory, will lead to less accurate conviction/acquittal decisions, showing in lower earnings. Although there is a difference in earnings between categories (average decision error is 54, 65 and 80 cents for category 1, 2 and 3, respectively), this difference is far from statistically significant (all p values are larger than 0.24 when tested on an individual level[16]). Apparently, participants in category 2 and 3 understand the nature of the evidence and make reasonable decisions,

but they do not reach these decisions in the manner normative theory supposes: first deriving a subjective probability of guilt and then making the decision.

We find generally no differences in the weights attached to incriminating and exonerating evidence.

Focusing on the participants, the behavior of whom is consistent with normative theory in the sense that they formed a reasonable belief about the guilt of defendants and therefore can be analyzed further, the general picture is as follows. They underestimate the strength of the evidence, and their subjective probability of guilt is biased in the direction of 50%. For the important group of cases with evidence pointing in the direction of guilt (combined strength of evidence >1), this means an underestimation of guilt. If the participants would act as value-maximizers, too few convictions would result. However, on average the cutoff point is lower than 80%, offsetting this effect. The net result is that too many suspects are convicted. In a clear breach of rationality, participants seem to be drawn towards the population rate of guilt, knowing that in the experiment 50% of the defendants are guilty. Another explanation could be probability matching[17] (see e.g. Shanks et al., 2002), which would imply that participants convict defendants with a probability equal to the subjective probability of guilt. This is, however, not supported by the data, given the consistent thresholds for conviction (cut-off points) we found for all participants (except one) who assess the probability of guilt systematically.

An alternative explanation is that participants are not confident about their assessment of the probability of guilt and that this would lead them not to calculate and use the correct threshold for conviction. We ran an additional treatment to examine this hypothesis[18]. For each of the thirty cases 43 participants[19] were asked to estimate the probability of guilt and were then given the objective probability, after which they had to decide between guilt and innocence. Note that this design makes it possible to learn over time which is not possible in the original design. Figure 5 maps the proportion of convictions against the objective probability guilt. Comparison with Figure 2 makes it immediately clear that decisions are closer to the optimum, but participants still convict defendants at much too low probability of guilt. 13 of the 43 participants applied the correct threshold. On average the threshold in terms of the objective probability of guilt was 71%. Participants were very consistent in using thresholds (27 never deviated form

their thresholds, while 14 deviated once; 2 participants were less consistent with 3 or 4 deviations).

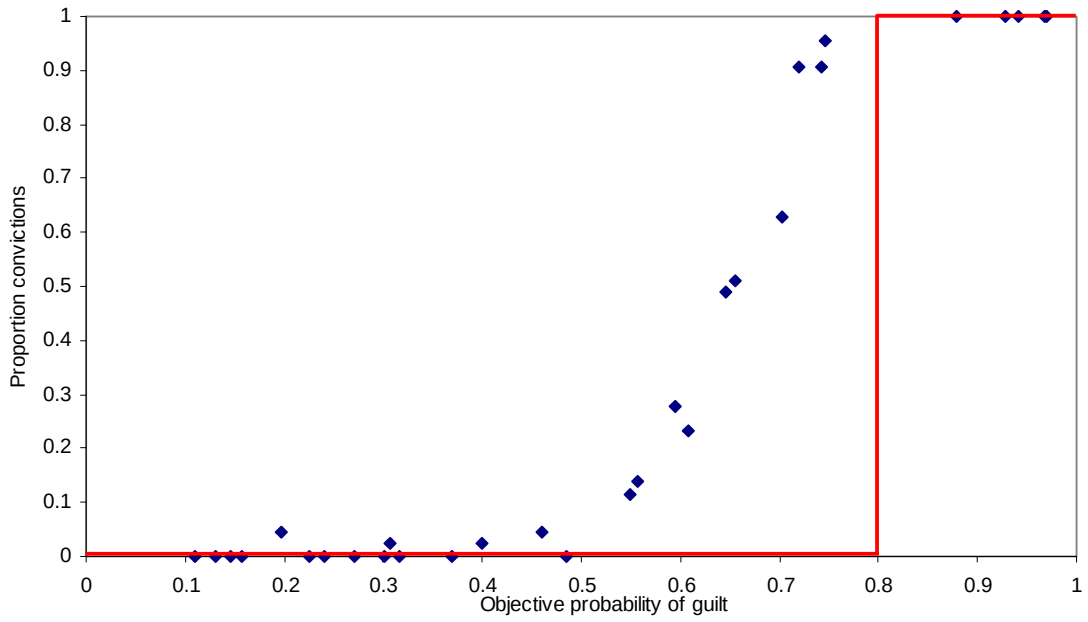


**Figure 5.** Proportion of conviction for the 30 cases, when participants are informed about the objective probability of guilt in each case. The red line shows the optimal decisions; when the objective probability of guilt is lower (higher) than .8 the defendant should be acquitted (convicted).

Like in table 6, we calculated the errors in the assessment of the probability and decision errors in terms of costs. It should be stressed that error in the assessment of probability can be measured as we asked participants to give this subjective assessment, but, as they could, and actually did, use the objective probability to decide the cases, subjective probability is not relevant the actual decisions. Because of the learning possibilities in this treatment, the quality of subjective probability does improve somewhat (the average error was 2.1 for category 1 participants[20] , $p<0.01$ Mann Whitney 2-sided test). The effect on the decision quality is much larger: the average decision error has been halved to 24.4 cents ($p<0.001$, Mann Whitney 2-sided test). To conclude, participants apply the rational model more consistently when they are provided with the objective probability of guilt than in the absence of such information. Confidence may therefore play a role.

However, the essential finding that participants convict at too low probability of guilt is reinforced by this additional session.

**6.2 Results part 2: inquiry and verdict**

The analysis is confined to decisions and errors, with particular emphasis on search behavior as potential source of error. We will, however, also look at the impact of background.

The optimal decision maker would convict in 40% of the cases, and our participants do so in 40.7% of the cases. This means that there is no general tendency to convict too few or too many suspects. However, the optimal decision maker would use on average 5.3 pieces of evidence, but the participants use on average only 4.4. This means that on average the participants have a distorted view of the probability of guilt, and this has severe consequences for the accuracy of their decisions and, thus, for their earnings. The expected earnings per case for the optimal decision maker are about 67 cents in this part (including the fixed salary per case), while average expected earnings of the participants are only 17 cents. The variance in earnings is enormous: 23% of the participants had negative expected earnings[21].

There are two kinds of errors participants can make. (1) They can gather too little or too much evidence and (2) they can make the wrong decisions given the evidence they gathered. It is possible that these two errors (partly) cancel each other.

First, we compare the evidence collected with the optimal amount of evidence (see figure 6). As the figure shows, in the majority of cases the participants stopped searching too soon (53%), in about 29% of the cases they stopped exactly at the optimal amount of evidence, and in 18% of the cases too much evidence was gathered. The tendency to search too little is in line with experimental research on sequential search, as noted before (see the references in Sonnemans, 1998).

Figure 6 also shows the decisions. When the right amount of evidence was gathered, the decision is almost always correct (95% of the cases). In the other cases we look at the correctness of the decision in the light of the available evidence. The decision is correct in 88% of the cases when too much evidence is gathered; of the errors that are made 59% are incorrect acquittals and 41% incorrect convictions. This suggests that these participants want to be on the safe side, gather more information than needed and

acquit relatively often. However, we did not find a relation between the risk attitude (measured in the first experiment; see section 5) and the amount of evidence gathered or the number of convictions. Most errors and the most serious ones were made when participants asked for too little evidence. Only in 74 percent of these cases the right decision was made, and on top of that the errors are biased towards unfounded conviction (56% of the errors).

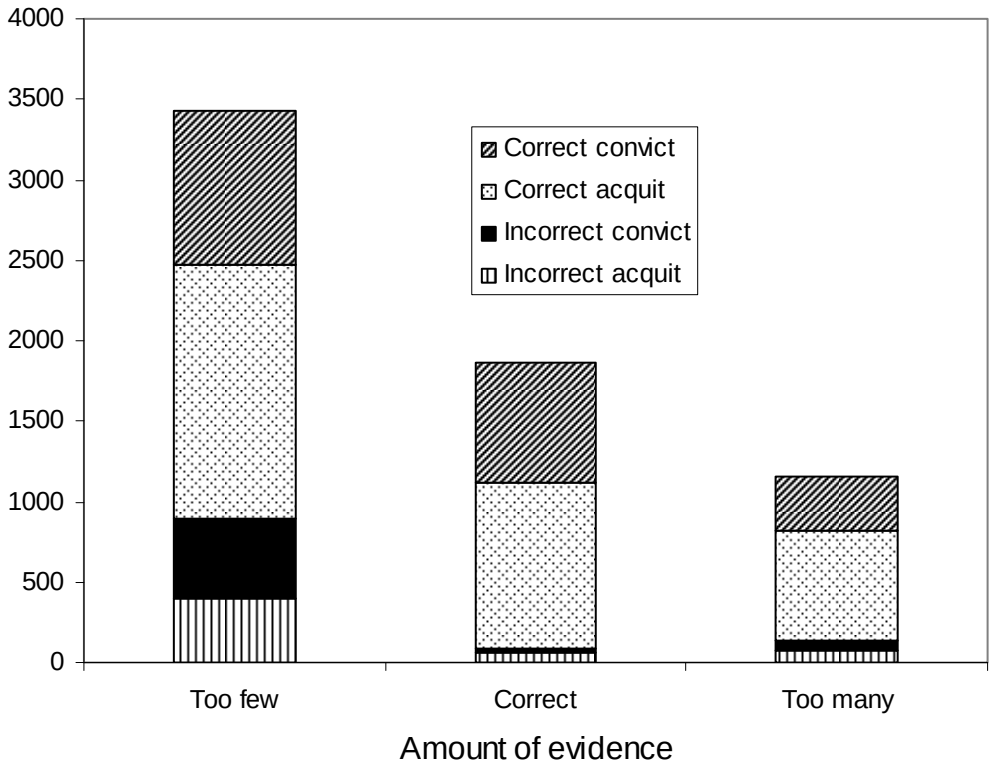


**Figure 6**: Frequency of decisions in part 2: the amount of evidence gathered by the participant and the decision to acquit/convict. The correctness of the decision is based upon the actually *gathered* evidence.

Note that there are large differences in the financial consequences of errors. To stop after two exonerating pieces of evidence and acquit the suspect is an error that will cost in expectation only 2 cents, while for example stopping after two incriminating pieces of evidence and convicting the suspect forms an error that will cost in expectation about 30 cents. To disentangle the financial consequences of errors, we calculate the *evidence error* as the expected earnings of the optimal strategy minus the expected earnings of the

optimal decision given the actually gathered evidence. The *decision error* is the difference between the optimal decision given the actually gathered evidence and the expected earnings of the actual decision. Note that the evidence error can be a negative number, for example when in a certain case it is ex ante *in expectation* optimal to search longer but ex post the extra evidence proves not to be very informative. The sum of these errors is the total error. Table 7 shows these errors for the different types of participants. We find that the evidence and decision error have about the same size. Although the decision and total errors of the law students seem larger than those of the science students, the differences are not statistically significant (tested on an individual level). The candidate judges do not differ from the student groups.

| **Subjects** | | **Evidence error** | **Decision error** | **Total error** | **N** |
|---|---|---|---|---|---|
| **Candidate judges** | | 24.83 | 20.60 | 45.42 | 53 |
| **Students** | Law | 22.49 | 29.95 | 52.44 | 51 |
| | Science | 21.32 | 20.70 | 42.03 | 25 |
| | Economics | 22.76 | 24.00 | 46.76 | 54 |
| | Other social science | 29.54 | 27.56 | 57.11 | 32 |
| Total | | 23.79 | 26.07 | 49.86 | 162 |

**Table 7**: Average errors in cents per case.

6.2.1 Discussion part 2

As expected, we find a tendency to search too little. In addition, 17.4% of the decisions were wrong given the available evidence. This number is surprisingly high when compared with the percentage wrong decisions in part 1 (16.8%). If participants found a decision hard, in part 2 they had the chance to order further inquiries up to the maximum of 7, and therefore one would expect fewer errors. However, the decision errors made in part 2 are on average less serious and thus less costly than in part 1, because unfounded acquittals and convictions are about equally represented (420 and 412, respectively) while in part 1 most errors were incorrect convictions. This reduces the average costs of decision errors by more than half.

We examined the causes of suboptimal search further by lowering the costs of evidence to two points[22]. The initial costs were 10 points, which is already low when compared with expected return in most situations[23]. Lowering the costs increases the optimal number of pieces of evidence from on average 5.3 to 5.7. We find that the participants are clearly sensitive to the cost of evidence: the actual number of requests increases from 4.4 to 5.5, much closer to the optimum. This, however, obscures large differences in behavior. In 40% of cases the participants stopped searching exactly at the optimal amount of evidence, in 29% the participants stopped too soon and in 31% too much evidence was gathered. At this low level of costs two types of behavior are observed. The first type is sequential search, and as before participants gather too little information on average. In the second type participants (nearly) always gather all evidence and then decide to convict or acquit[24]. These participants searched too long on average. Again most errors were made when decisions were based on too little evidence, followed by decisions based on too much evidence.

## 7. Conclusion

We formulated six research questions in section 4, and addressed them in the experiment. To summarize the answers:

1. When evidence is inherently uncertain, as in reality is always the case, and evidence needs to be combined, which is also the case in most legal systems, verdicts are inaccurate (see figure 2). This occurs, despite the important finding that all relationships between evidence and verdict that are displayed in the decisions of the participants are qualitatively correct (correct sign, correct order, etc.). It is particularly worrisome that, for given evidence (part 1), errors are biased towards the most serious type: unfounded conviction. In clear cut cases, that presumably dominate in actual court rooms, the inaccuracy will not show, but in complicated cases, in which evidence is relatively weak and/or contradictory, error will be abundant. On the positive side, there is no tendency to give different weight to incriminating and exonerating evidence.
2. Inaccuracy has two causes: half of the participants do not use the rational approach, embodied in the normative model. They do not asses the probability of

guilt quantitatively, and thus cannot base their verdicts on it. The other half does act in ways (roughly) consistent with rationality, but does this in a very imprecise manner. Decision error does not differ significantly among these groups.

3. In line with earlier research (see section 2), participants, in as far as they form beliefs about probability of guilt, underestimate the strength of (both incriminating and exonerating) evidence. Their assessment of the probability is too close to 50%. For the relevant cases this phenomenon would lead to too many acquittals. Again, we find in general no bias against exonerating evidence.
4. However, these participants systematically convict defendants at a probability of guilt that is too low. This more than compensates the underestimation of probability. Thus, the prevalence of unfounded convictions noted above is caused by making wrong decisions, given subjective probability. Participants seem to be guided too much by the prior, knowing in the experiment that 50% of the defendants were guilty.
5. When participants have the possibility to search for evidence, we find, as expected, that they do not search long enough. Still, the other potential source of error, making wrong decisions given the evidence, is roughly as important. An important difference with part 1, is, however, that errors are less severe. The numbers of unfounded acquittals and convictions are roughly the same.
6. With respect to the impact of background, candidate judges and law students performed worse in part 1 of the experiment. The difference was not caused by the assessment of probability, but by the verdicts, given subjective probability. In part 2 differences were not statistically significant.

In this experiment we looked at the task of judges in a very limited way. Their work is much richer, and requires much more abilities than statistical reasoning. Nonetheless, uncertainty is so fundamental to adjudication of criminal cases that without the competent handling of uncertainty these other abilities lose much of their relevance in hard cases. The study shows that, although the participants, candidate judges and students in different disciplines including top law students, understand the basics, most of them even among these candidate judges and top law students lack the skills to reach correct verdicts, when

it comes to hard cases. It seems safe to conclude that even talented and highly educated people need a lot of training to make correct decisions under uncertainty. We have no reason to assume that acting judges will perform much better than our participants, noting that judges are recruited from this group. We find that many participants make decisions in an intuitive way: half of the participants do not reason systematically from probability of guilt to verdict; the others do reason systematically, but in a very imprecise way and seem to be led astray by their strong focus on the prior probability of guilt. Reliance on intuition makes it difficult to comprehend and verify verdicts, and this is increasingly likely to erode the confidence of (forensic) experts and the general public in verdicts. We, therefore, agree with Guthrie et al. (2007) that rational deliberation needs to take a more prominent place in court rooms.

Our experiment is unrealistic in the sense that in order to compare actual and optimal decisions quantitative information about all evidence was made available. In judicial practice this is not the case, although, as we noted in the introduction, the availability of quantitative information is rapidly increasing and we expect this trend to continue. It seems unavoidable that due to the much less structured nature of information in real situations verdicts in hard cases are even more prone to error than our experiment would suggest. At least in hard cases we feel strongly that the qualitative tradition of legal decision making needs to be replaced by a quantitative, probability based approach. Hard cases can be broadly demarcated as all cases without multiple, consistent and independent evidence for guilt (or innocence), or more narrowly as a range of the probability of guilt.

The result that for given evidence most errors are unwarranted convictions requires specific attention. Our experiment (above, point 4) suggests that within the legal framework of given evidence the remedy should focus on the probability of guilt minimally needed to convict a defendant. The current qualitative standards such as ‘beyond reasonable doubt’ give little guidance, and a sufficiently high quantitative standard would help to reduce error. More fundamentally, allowing judges to order further enquiries would take away the bias towards unwarranted convictions (above, point 5).

**Literature**

Allen, Ronald J. 2007. "The problematic value of mathematical models of evidence," 36 Journal of Legal Studies 107-139.

Bornstein, Brian H. 2004. "The impact of different types of expert scientific testimony on mock jurors' liability verdicts," 10 Psychology, Crime and Law 429-446.

Brink, Axel, Lambert Schomaker and Marius Bulacu. 2007. "Towards explainable writer verification and identification using vantage writers," ICDAR 824-828.

Buchanan, Mark. 2007. "Conviction by numbers," 445 Nature 254-255.

Clark, Steven E. and Gary L. Wells. 2007. "On the diagnosticity of multiple-witness identification," Law and Human Behavior Online First.

Dawd, Philip. 2005. "Statistics on trial," 2 Significance 6-8.

Donnelly, Peter. 2005. "Appealing statistics," 2 Significance 46-48.

Eddy, David M. 1982. "Probabilistic reasoning in clinical medicine: problems and opportunities," in D. Kahneman, P. Slovic and A. Tversky, eds., Judgment under uncertainty: heuristics and biases. Cambridge, UK: Cambridge University Press.

Edwards, Ward. 1968. "Conservatism in human information processing", in B. Kleinmuntz, ed., Formal representation of human judgment. New York: Wiley.

Emons, Winand and Claude Fluet. 2009. "Accuracy versus falsification costs: the optimal amount of evidence under different procedures," 25 Journal of Law, Economics and Organization 134-156.

Faigman, David L. and A.J. Baglioni. 1988. "Bayes' theorem in the trial process: instructing jurors on the value of statistical evidence," 12 Law and Human Behavior 1-17.

Froeb, Luke M. and Bruce H. Kobayashi. 1996. "Naive, biased, yet Bayesian: can juries interpret selectively produced evidence?" 12 Journal of Law, Economics and Organization 257-275,

Guthrie, Chris, Jeffrey J. Rachlinski and Andrew J. Wistrich. 2001. "Inside the judicial mind," 93 Cornell Law Review 1-43.

Guthrie, Chris, Jeffrey J. Rachlinski and Andrew J. Wistrich. 2007. "Blinking on the bench: how judges decide cases," 86 Cornell Law Review 777-830.

Holt, Charles A. and Susan K. Laury. 2002. "Risk aversion and incentive effects", 92 American Economic Review 1644-1655.

Jonakait, Randolph N. 1983. “When blood is their argument: Probabilities in criminal cases, genetic markers, and once again Bayes' theorem,” University of Illinois Law Review 369- 421.

Kahneman, Daniel and Amos Tversky. 1972. “Subjective probability: A judgment of representativeness,” 3 Cognitive Psychology 430-454.

Kassin, Saul M., Richard A. Leo, Christian A. Meissner, Kimberly D. Richman, Lori H. Colwell, Amy M. Leach and Dana LaFon. 2007. “Police interviewing and interrogation: A self-report survey of police practices and beliefs,” 31 Law & Human Behavior 381-400.

Koehler, Jonathan J. 1996. “The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges,” 19 Behavioral and Brain Sciences 1-53.

Landes, William M. and Richard A. Posner. 2001. “Harmless error,” 30 Journal of Legal Studies 161-192.

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

Pratt, John W., Howard Raiffa and Robert Schlaifer. 1995. “Statistical decision theory,” Cambridge, Mass: MIT Press.

MacCoun, Robert J. 1999. “Epistemological dilemmas in the assessment of legal decision making,” 23 Law and Human Behavior 723-730.

Martin, Anne W. and David A. Schum. 1987. “Quantifying burdens of proof: a likelihood ratio approach,” 27 Jurimetrics Journal 383-402.

McAllister, Hunter A. And Norman J. Bregman. 1986. “Juror underutilization of eyewitness nonidentifications: theoretical and practical implications,” 71 Journal of Applied Psychology 168-170.

Meester, Ronald, Marieke Collins, Richard Gill and Michiel van Lambalgen. 2006. “On the (ab)use of statistics in the legal case against the nurse Lucia de B.,” 5 Law, Probability and Risk 233-250.

Mood, A.M. , F.A. Graybill and D.C. Boes (1973). Introduction to the theory of statistics. New York: McGraw- Hill Kogakusha.

Offerman, Theo, Joep Sonnemans, Gijs van de Kuilen and Peter P. Wakker. 2009. "A Truth-Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes," 76 Review of Economic Studies 1461-1489.

Robinson, Lori B. and Reid Hastie, 1985. Revision of Beliefs When a Hypothesis Is Eliminated From Consideration. 11 Journal of Experimental Psychology: Human Perception & Performance 443-456.

Saks, Michael J. and Robert F. Kid. 1980. "Human information processing and adjudication: trial by heuristics," 15 Law and Society Review 123-160.

Schrag, Joel and Suzanne Scotchmer. 1994. "Crime and prejudice: the use of character evidence in criminal trials," 10 Journal of Law, Economics and Organization 319-342.

Schum, David A. and Anne W. Martin. 1982. "Formal and empirical research on cascaded inference in jurisprudence," 17 Law & Society Review 105-151.

Shanks, David R., Richard J. Tunney and John D. McCarthy. 2002. "A re-examination of probability matching and rational choice," 15 Journal of Behavioral Decision Making 233-250.

Sonnemans, Joep. 1998. "Strategies of Search," 35 Journal of Economic Behavior and Organization 309-332.

Sonnemans, Joep. 2000. "Decisions and Strategies in a Sequential Search Experiment," 21 Journal of Economic Psychology 91-102.

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

Thompson, William C. and Edward L. Schumann. 1987. "Interpretation of statistical evidence in criminal trials," 11 Law and Human Behavior 167-187.

Tversky, Amos and Daniel Kahneman. 1992. "Advances in prospect theory: Cumulative representation of uncertainty," 5 Journal of Risk and Uncertainty 297-323.

Van der Heide, Wieger, Frank van Tulder and Caspar Wiebrens. 2007. "Strafrechter en strafketen: de gang van de zaken, 1995-2006," 3 Rechtstreeks 7-72.

Van Koppen, Peter J. 2008. "Blundering justice: The Schiedam Park Murder," in R.N. Kocsis, Serial murder and the psychology of violent crimes, New York: Humana.

Wagenaar, Willem A. 2006. "Vincent plast op de grond; nachtmerries in het Nederlandse recht," Amsterdam: Bert Bakker.

Way, Vicky. 2003. "Judicial fact-finding by judges alone in serious criminal cases," 27 Melbourne University Law Review 423-457.

---

[1] See e.g. Froeb and Kobayashi (1996) and Emons and Fluet (2007) who both model the selective presentation of evidence by parties.

[2] See also the claim of Koehler (1996) that the position that people routinely ignore base rates has been vastly overstated.

[3] Other errors may be committed as well, for instance, procedural mistakes.

[4] The Netherlands justice system was recently shaken by such a case. In the so called Schiedam park murder case it became clear by finding unmistakably the real perpetrator that the wrong person had been convicted of a child murder. See Van Koppen, 2008. In another high profile case which involved complicated statistics it was also concluded that the conviction was unwarranted (Meester et al., 2006).

[5] Note that in many instances it is not straightforward which objective assessment to apply (see Clark and Wells, 2007, Koehler, 1996 and Allen and Pardo, 2007). We will ignore this problem here.

[6] Or, dependant on the legal system, the jury. Jury decision making is not part of this study.

[7] Remember that the participants learned only after the last case of part 2 which of the accused were guilty and how much their earnings were. This prevented that participants with negative earnings would take extra risk to get a positive balance again.

[8] If in each case both belief and decision are rewarded, participants may be tempted to report a high belief of guilt, but acquit the accused, as one of these would have a positive pay-off.

[9] For later reference, we denote the 34 participants of whom the response could not be used as category 3.

[10] In principle this could be an effect of the quadratic scoring rule that is used because this rule is only incentive compatible for risk neutral decision makers and extreme risk-

aversion participants should report probabilities closer to 50% (Offerman et al 2008). We have for each participant a measure of risk-aversion and we find no significant relation between the standard deviation of the probabilities the participant reported and the risk-aversion (Pearson correlation is 0.06, rank correlation 0.04).

[11] The low cutoff points could be caused by risk aversion. However, we find no significant correlation between risk-aversion (as measured before the experiment in the first sessions) and cutoff point. Alternatively we can calculate a cutoff point using our measures of risk and loss aversion. Assuming a utility function $U(x)= x^{\rho}$ for $x>0$ and $U(x)=\lambda x^{\rho}$ for $x<0$ the cutoff point is $(\lambda 1500^{\rho}+ 100^{\rho})/(2* 100^{\rho}+ \lambda 300^{\rho}+ \lambda 1500^{\rho})$. Although most (two thirds) of these points are in the range 70-90 and look reasonable, they do not correlate at all with the fitted cutoff point described above. By construction, the fitted cutoff point is more in line with the data. We did not measure the risk attitude in later sessions with law students and candidate judges because of a time constraint.

[12] This line is practically the same for all three categories of participants and therefore the groups are aggregated.

[13] If we compare the separate groups of law students and candidate judges, we find that candidate judges perform worse than science students (Mann Whitney 2-sided test $p=.06$) and that law students perform worse than other students (2-sided Mann Whitney test $p<0.05$, $p<0.01$ and $p<0.01$ for comparison with science, economics and other social sciences respectively)

[14] Mann Whitney 2-sided tests, no differences between law-science and economics-other, $p<0.001$ for comparisons judges-economics, judges-other, law-economics, law-other, science-economics, $p<.05$ for science-other and candidate judges versus law students.

[15] Including category 2 subjects increases the errors but does not lead to differences between disciplines of subjects.

[16] If the tests are performed on the level of the decisions (ignoring that decisions by the same decision maker are not independent), the errors are smaller for category 1 than the other categories (Mann Whitney 2-sided tests both p's $<0.05$).

[17] As pointed out by an anonymous referee.

[18] We thank an anonymous referee for this suggestion.

[19] Participants were students in economics (20), other social sciences (15) and science students (8); no law students were involved.

[20] Again, we divided the participants in categories based on individual logistic regression. The large majority (37) is of category 1, 6 participants are in category 2 and we find no participants in category 3.

[21] When realized earnings in part 2 were negative the participant earned 0 in this part but kept their earnings of part 1. Note that the participants learned only after the final case of part 2 whether the suspects were guilty or not and also their earnings.

[22] We thank a referee for this suggestion.

[23] For example, after two pieces of incriminating evidence (and no exonerating evidence) the information value of an extra piece of evidence is about 40 cents, which is much more than the costs of 10 cents.

[24] Participants who make their decisions in this way, will sometimes make a decision counter to the last piece of evidence (last decision was exonerating and they convict, or last evidence was incriminating and they acquit). This happens in 18.4% of all decisions compared with 11.3% in the original treatment.