ERRORS IN JUDICIAL DECISIONS: EXPERIMENTAL RESULTS

Joep Sonnemans[a] and Frans van Dijk[b]

[a] CREED, Amsterdam School of Economics, Universiteit van Amsterdam, Roetersstraat 11, 1018WB Amsterdam, The Netherlands and the Tinbergen Institute.
Email: J.H.Sonnemans@uva.nl

[b] Council for the Judiciary, Kneuterdijk 1, 2509 LP The Hague, The Netherlands.
Email: F.van.Dijk@rechtspraak.nl

September 2008

Abstract

In criminal cases the task of the judge is to transform the uncertainty about the facts into the certainty of the verdict. In this experiment we examine the relationship between evidence of which the strength is known, subjective probability of guilt and verdict for abstract cases. We look at two situations: (1) all evidence is given and (2) evidence can be acquired. Roughly half of the participants do not base their decision on a subjective belief of the probability of guilt. The others underestimate in general the probability of guilt, but this is more than compensated by a tendency to convict at too low probability of guilt. In the situation where evidence can be acquired, participants do not acquire enough evidence.

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'. When all evidence points in the same direction, verdict is easily reached. Hard decisions have to be taken, when evidence is weak and/or contradictory.

From a decision-theoretic perspective (see e.g. Pratt et al., 1995), judicial decision making is just a case of decision making under uncertainty. 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 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. They link these biases with reliance on intuition instead of rational deliberation, as taking recourse to intuition can lead easily astray when dealing with probability. They show that intuition plays an important role in judicial decision making. In their view, judicial accuracy is, consequently, the most challenging issue facing the courts.

The importance of this issue is increasing, as challenges to incorrect handling of uncertainty associated with evidence are mounting. 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, forthcoming, Kassin et al., 2007). Consequently, judges can dispose of more precise information, which lends itself or even requires to be combined by mathematical means. Judges are confronted with or bring to court more and more experts who provide such precise information. What is also becoming increasingly evident, is that the incorrect interpretation or combination of evidence makes judges vulnerable to technical criticism, either by experts who have provided testimony in specific cases and who believe that the evidence they provided has been misinterpreted in verdicts, but also by members of the public, in particular scientists, who - often passionately - believe that a miscarriage of justice has occurred. See Meester et al. (2006) and Buchanan (2007) with respect to one in a series of such cases in the Netherlands. Explaining that the judge used his intuition is obviously not the right answer to these critics if demonstrably mistakes have been made in interpreting and combining probabilities, even if the verdicts as such may not be wrong.

The importance of judicial accuracy raises the question with what precision people are actually able to handle probabilistic decision problems of the type judges are confronted with, and, consequently, to what extent errors are made. This is a complex matter and we focus exclusively on the interpretation and combination of evidence of which the strength (diagnostic value) is completely known. We will try to answer the question to what extent people measure up to a normative standard, i.e. 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.

To focus on participants' handling of uncertainty, we address these questions in an experiment that captures key aspects of judicial decision making in an abstract manner. Descriptive information about cases and evidence is eliminated as much as possible. 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 accused are guilty. In the second part only some 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).

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 entailed 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 guilt, *ng* for not guilty and *e* for evidence:

$$P(g|e)/P(ng|e) = P(e|g)/P(e|ng)*P(g)/P(ng) \quad (1)$$

where $P(g|e) + P(ng|e) = 1$

The ratio $P(e|g)/P(e|ng)$ expresses the strength of the evidence or its diagnostic value. If this ratio is larger than one, the evidence is incriminating. If the ratio is smaller than one, the evidence is exonerating. Using the equality, $P(g|e)$ can be expressed in the terms of the right hand side.

The literature shows that most people have difficulty thinking in probabilities. They have problems with assessing the individual probabilities of the right hand side, and they have trouble combining these probabilities. The problems of aggregation are of interest here. In several experiments it has been found that compared to this standard participants do not give sufficient weight to the evidence. In one form or the other, participants were given or first asked to assess the right hand side probabilities and then 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 and let them assess the probability of guilt again. 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). It seems reasonably well established that participants tend to underutilize 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 shown 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 as example the demeanor of the defendant. 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: $P(g|e) = P(e|g)$. This fallacy is also documented, for example, 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 a multitude of examples of such bias is given. In catchy words, Thompson and Schumann (1987) call it the Prosecutor's fallacy, i.e. overvaluing evidence by not taking into account the a priori likelihood that a defendant is not guilty. However, they also find an opposite fallacy, which they term the Defense Attorney's Fallacy and which means that probabilistic evidence is completely ignored. This happens in particular when probabilities are not expressed in terms of percentages, but in terms of matches in a relevant population. They

find that sizable numbers of participants fall prey to both fallacies, but that the Defense Attorney's Fallacy, i.e. the non-utilization of evidence, is the more serious problem. 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 bias, this is not the case for the whole group that does not disregard the evidence. These findings fuel the claim of Koehler (1996) that the position that people routinely ignore base rates has been vastly overstated.

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, which is of particular interest here. 41% was not subject to the representativeness 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 overutilized the evidence. It should be noted, however, that this study was not a controlled experiment, but more like an intellectual exercise.

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. These findings link 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. They show that all witness responses should be taken into account and not only that of the witness(es) who identified the suspect, while ignoring witnesses who did not. 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) distinguishes 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 not conclusive. The research suggests sources of error, the importance of which is, however, not consistently established. Also, the research does not address the issue whether the sources of error would actually lead to errors. While wrong assessments of probability of guilt given the evidence may be made, it is not established whether in fact this leads to wrong decisions. For instance, while people may underestimate the probability of guilt, they may convict at a lower probability threshold than rationality would require, given standards and preferences. The reverse may also be true. To address these issues a more integral approach which takes probability assessments and decisions into account is needed of judicial decision making.

3. Conceptualization

Errors and incentives

From the perspective of accuracy of judicial decisions, judges can make two types of error[1]:

[1] From other perspectives errors may be committed as well, for instance, procedural mistakes.

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 can of course not be ascertained independently. Only in exceptional cases convicts are unequivocally exonerated, because the real perpetrator turns up[2] 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 Schiedam Park murder case in the Netherlands, mentioned in footnote 2, 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).

[2] The Netherlands justice system was recently shaken by such a case. In the so called Schiedam park murder case it became clear by finding the real perpetrator unmistakably that the wrong person had been convicted of a child murder. See Van Koppen, 2008.

		Real situation the accused is the perpetrator	Real situation the accused is 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.

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 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:

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

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 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 much as possible, objective assessment of this strength, for instance based on scientific research[3]. Basic beliefs 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. In the experiment the prior has to be specified and given to the participants.

Decision and search

When all allowable evidence is presented by prosecution and defense, the task of the judge (or jury, of course) is to decide the case. This requires him, however qualitatively and intuitively, to evaluate his subjective assessment of guilt against a threshold for conviction. This threshold depends on his incentives, as discussed above, within the legal framework ("beyond reasonable doubt" or "convincingly proven"). For the justice system as a whole, fair trial would necessitate that there is sufficient uniformity across cases and judges.

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 (for the latter see Way, 2003). Prosecution

[3] Note that in many instances it is not straightforward which objective assessment to apply (see Clark and Wells, 2007, and Koehler, 1996.

and defense present evidence, but the judge questions (expert) witnesses, decides whether or not 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 against specific costs.

4. Research questions

We can now formulate the research questions to be answered by the experiment. *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 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 do their verdicts differ 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?

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 Laurie, 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[4].

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. 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 of a single type could take place.

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.

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 discussed 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 presented was sorted by kind (incriminating or exonerating) and strength.

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 belief (subjective assessment of the probability of guilt) was rewarded (both with probability 50%): 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. This scoring rule is incentive compatible for risk neutral individuals (see Offerman et al., 2008). This procedure prevents hedging behavior by participants[5]. All participants received the same cases and evidence.

[4] 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.

[5] 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.

The risk neutral optimal decision maker is indifferent between conviction and acquittal when ap+b(1-p)=cp+d(1-p) with p the probability of guilt. With the chosen parameters this solves for p=0.8. 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.

Part 2: inquiry and verdict

Only one piece of evidence was given. And 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 an 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 a verdict guilty or not guilty. If he decided to order an inquiry, he, subsequently, had to decide on a further inquiry, and so on. 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 an accused 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 three groups: law, science and social sciences. The last group consisted of economics and psychology students. Most law students who participated were so called "honors' students". These students are the top 10% of their year, but we consider this not a validity threat because most judges are also recruited from the top segment. To facilitate participation the experiment took place at the regular Creed laboratory and in a computer room of the law school.

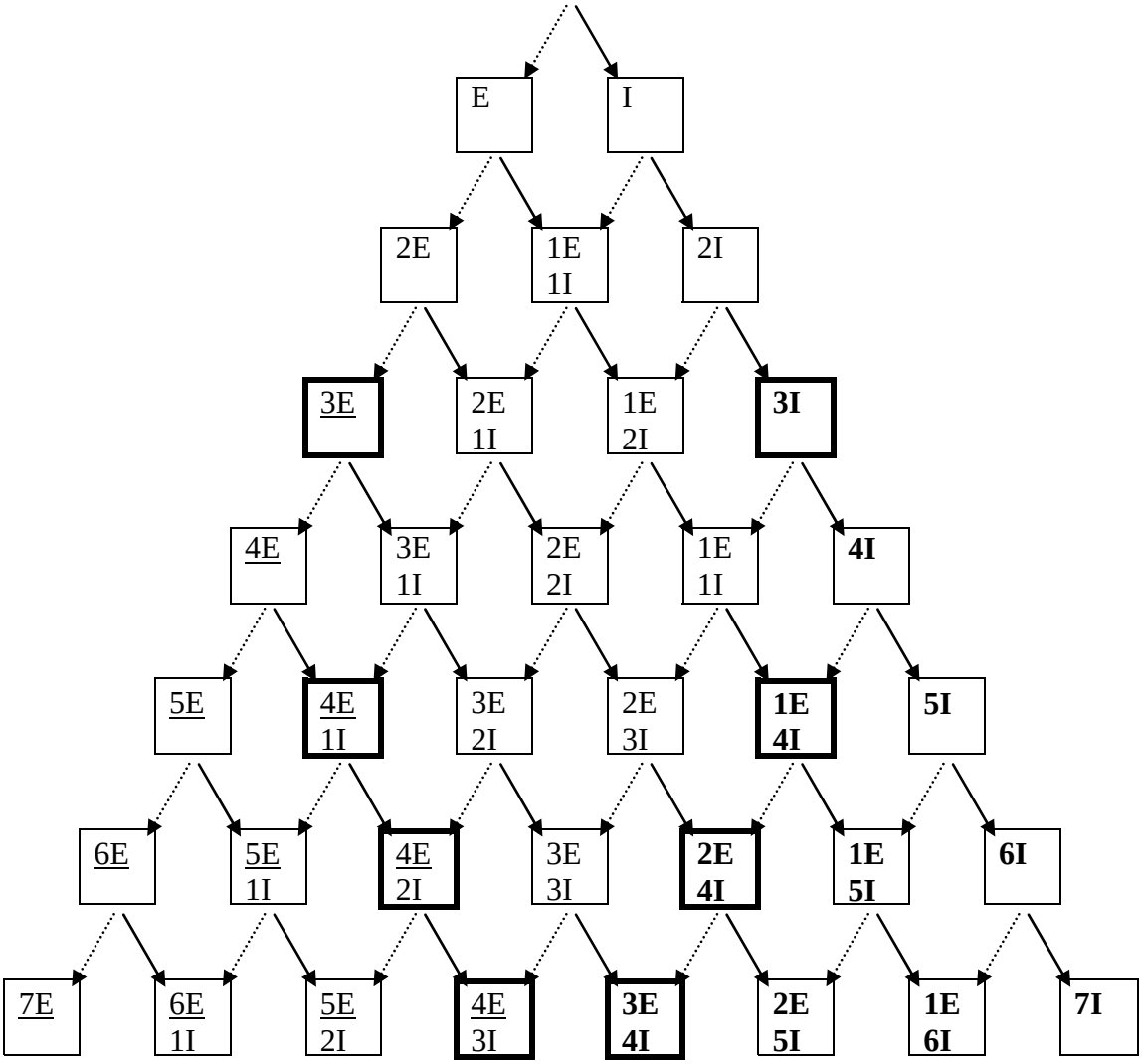

Figure 1. Outcomes and optimal strategies in the search part.

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 analysed. In that subsection we will also look at the impact of differences in background of participants (law, science, economics and other social sciences).

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

$Z=-1.06+1.50*INC_1+2.19*INC_2+0.63*INC_3-2.40*EXO_1-1.48*EXO_2-0.58*EXO_3$

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 87.7% 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 -1.06 implies that a priori (or when incriminating and exonerating evidence have the same strength) the participants will convict in about 25% ($1/(1+e^{1.06})$) of the cases.

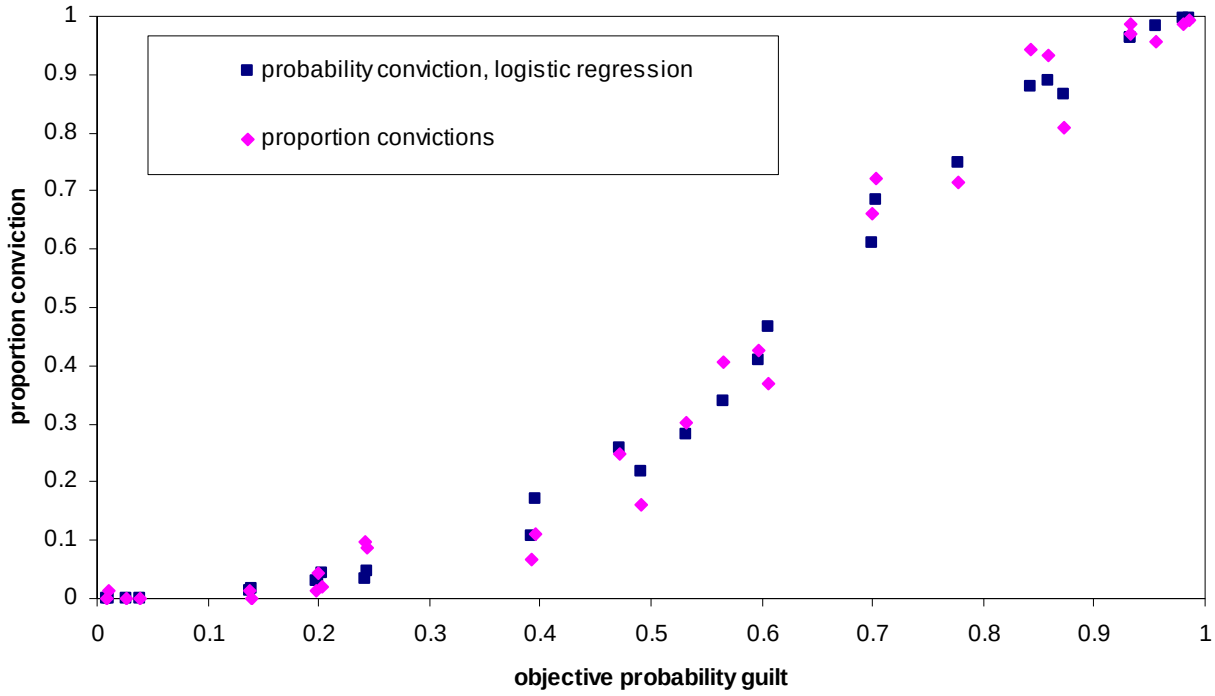

Figure 2. Proportion of conviction (diamonds) and the predicted proportion according to a logistic regression (squares) for the 30 cases.

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 neglected 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 different elements of evidence 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 subjective odds should be the product of the prior (1) and the odds of the evidence:

$$Odds_{Subj} = Odds_{\text{Prior}} * 2.33^{INC_1} * 4^{INC_2} * 1.5^{INC_3} * 0.25^{EXO_1} * 0.43^{EXO_2} * 0.67^{EXO_3}$$

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:

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

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 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 16 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 131 participants in two categories. Category 1 consists of the (82) 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 (49) 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 forms expectations[6].

We also calculated a combined regression for the 131 participants in category 1 and 2 together (top panel of table 4) and for the 82 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 significantly) than 1, meaning that a priori the participants consider the probability that the defendant is guilty

a little less than 50% but 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 prediction. It is given not enough weight. Theoretical neutral combinations of exonerating and incriminating evidence (total odds about 1) lead to subjective odds that are also close to 1, as long as 1EXO is not part of the evidence. For example, focusing on category 1, the two pieces of evidence 1EXO and 2INC should cancel each other exactly (odds are 0.25*4=1) but combine to subjective odds of 3.24*0.46==1.49, incriminating instead of neutral. For medium strong evidence (1INC and 2EXO) the combination has subjective odds of 2.02*0.55=1.11 which is only slightly larger than 1: marginally incriminating. The objective neutral combination of weak evidence 3EXO and 3INC combines to an odds of 1.35*0.68=0.92, slightly exonerating. The combination "3INC 3INC 2EXO" leads to subjective odds of 1.00 (objective odds 0.96) and combination "3EXO 3EXO 1INC" leads to subjective odds 0.93 (objective odds 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%.[7] The pattern for category 1 looks very similar to the probability weighting functions as used in prospect theory (Tversky and Kahneman,1992).

[6] For later reference, we denote the 31 participants of whom the response could not be used as category 3.

[7] 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).

	Categories 1 and 2 (N=131)						
Variable	*B*	*SE B*	*Beta*	*T*	*Sig T*	*Subjective odds* e^B	*Objective odds*
INC_1	0.66	0.03	0.32	21.60	0.00	1.94	2.33
INC_2	1.08	0.05	0.31	22.02	0.00	2.96	4
INC_3	0.27	0.03	0.12	9.26	0.00	1.30	1.5
EXO_1	-0.53	0.04	-0.21	-14.06	0.00	0.59	0.25
EXO_2	-0.36	0.03	-0.16	-11.49	0.00	0.69	0.43
EXO_3	-0.30	0.03	-0.12	-9.29	0.00	0.74	0.67
(Constant)	-0.22	0.10	-2.21	0.03		0.80	1

	Category 1 only (N=82)						
Variable	*B*	*SE B*	*Beta*	*T*	*Sig T*	*Subjective odds* e^B	*Objective odds*
INC_1	0.71	0.03	0.34	23.08	0.00	2.02	2.33
INC_2	1.18	0.05	0.33	23.95	0.00	3.24	4
INC_3	0.30	0.03	0.14	10.49	0.00	1.35	1.5
EXO_1	-0.77	0.04	-0.30	-20.58	0.00	0.46	0.25
EXO_2	-0.60	0.03	-0.26	-19.10	0.00	0.55	0.43
EXO_3	-0.38	0.03	-0.15	-11.97	0.00	0.68	0.67
(Constant)	-0.14	0.10	-1.40	0.16		0.87	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 131 participants (categories 1 and 2, see main text). Adjusted R-square is 0.47
Lower panel: data of the 82 participants (category 1 only, see main text). Adjusted R-square is 0.68.

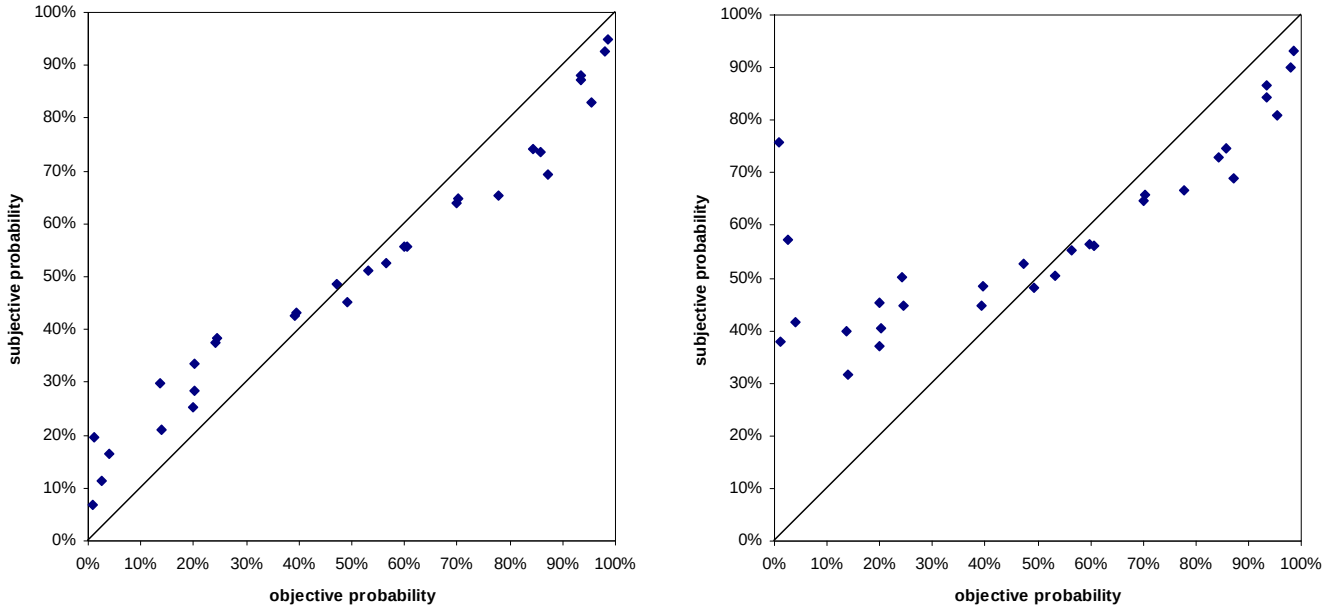

Figure 3: Scattergram of the average subjective probability for all 30 cases of part 1, with on the horizontal axis the objective probability. In the left panel the 82 participants of category 1, in the right panel the 49 participants of category 2. See the appendix for the same figure for these 31 participants.

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 steep 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 82 participants of category 1 are displayed in table 5. Exactly half of the participants (41) 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 63.15%, 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)[8]. If

[8] The low cutoff points could be caused by risk aversion. However, we find no significant correlation between risk-aversion (as measured before the experiment) and cutoff point. Alternatively we can calculate

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).

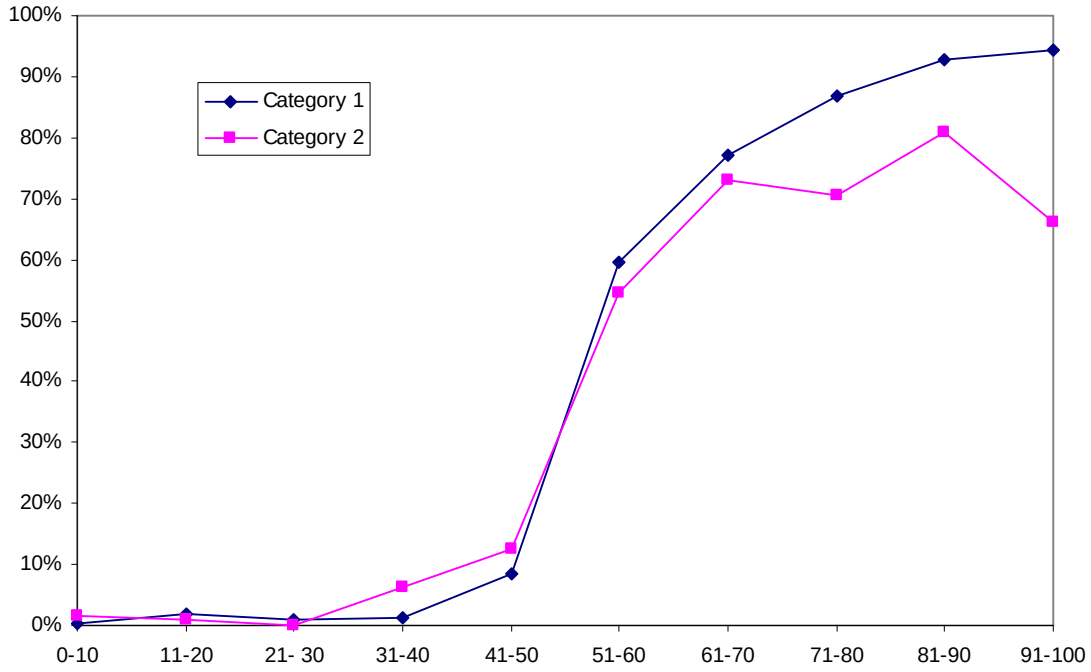

Figure 4: Average conviction rates by subjective probability of guilt. A separate line is drawn for participants in category 1 and 2.

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.

		Errors					
		0	1	2	3	7	Total
Cutoff	50	4	3	1	1		9
	51	1	1				2
	52	1	1				2
	53				1		1
	55	8	2	1			11
	58	1	1				2
	60	11	6	2		1	20
	65	2	4	2	1		9
	66	1					1
	67		1				1
	68		1				1
	69	1					1
	70	5	2	1			8
	75	2	3				5
	78	1					1
	80	3	2	1			6
	100		1		1		2
Total		41	28	8	4	1	82

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.

6.1.4. Errors and individual differences

In 83.7% of the cases the decision equals the optimal decision. Of the 793 deviations from optimality the majority is of the most serious kind: 725 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 94.8% 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 3.6% of the suspects are convicted. However, in the large area in between, where the evidence points in the direction of guilt, but is not strong enough to rationally convict the suspect, much too many suspects are convicted. As a result, in these 9 cases the participants lost 11.50 euro on average (which most 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 beliefs. As to the verdicts, 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). 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)[9].

	Assessment of probability		**Decision**	
Students	**Average Error**	**N**	**Average Error**	**N**
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	82	59.4	162

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 background. For the assessment of probability only participants of category 1 are included.[10]

6.1.5 Discussion part 1

Apparently, probability concepts do not come naturally to most individuals (as many professors in statistics can testify). About half of the participants (82) report consistently

[9] Mann Whitney 2-sided tests, $p<0.001$ for comparisons law-economics, law-other, science-economics and science-other and no differences between law-science and economics-other.

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, 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 50, 65 and 83 cents for category 1, 2 and 3, respectively), this difference is far from statistically significant (all p values are larger than 0.25 when tested on an individual level[11]). 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 no differences in the weights attached to incriminating and exonerating evidence.

Focusing on the participants whose behavior is consistent with normative theory, 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.

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 also look at the impact of background.

The optimal decision maker would convict in 40% of the cases and our participants do so in 39.8% 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

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

average 5.3 pieces of evidence, and our 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: 26% of the participants had negative expected earnings[12].

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 5). As the figure shows, in the majority of cases the participants stopped searching too soon (51%), in about 30% of the cases they stopped exactly at the optimal amount of evidence, and in 19% of the cases too much evidence was gathered. The tendency to search too little is in line with experimental research on sequential search (see the references in Sonnemans, 1998).

Figure 5 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 89% of the cases when too much evidence is gathered; of the errors that are made two thirds are incorrect acquittals and one third 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) 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 76 percent of these cases the right decision was made, and on top of that the errors are biased towards unfounded conviction (55% of the errors).

---

[11] 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).

[12] 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.

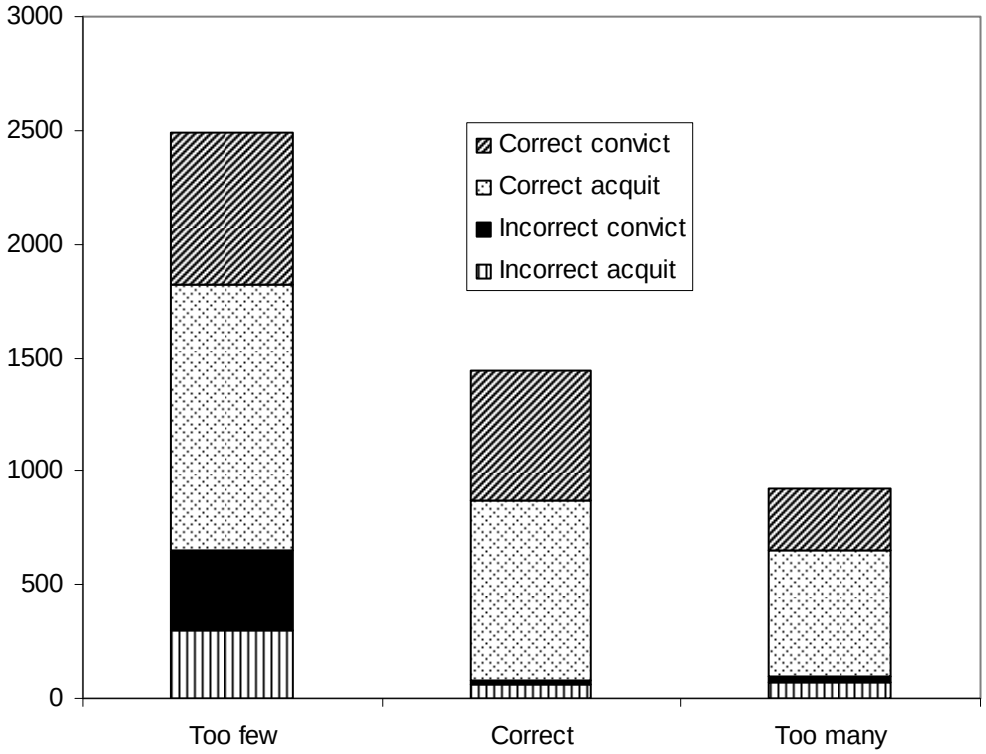

Figure 5: Frequency of decisions in part 2: the amount of evidence gathered by the participant and the decision to acquit/convict. The correctness of the decision is based upon the actually *gathered* evidence.

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).

Students	Evidence error	Decision error	Total error	N
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, 17.1% of the decisions were wrong given the available evidence. This number is surprisingly high when compared with the percentage wrong decisions in part 1 (16.3%). 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.

7. Conclusion

We formulated six research questions in section 3, and addressed them in the experiment. To summarize the answers:

1. When evidence is inherently uncertain, as in the real world is always the case, and evidence needs to be combined, which is the case in most legal systems, verdicts are inaccurate (see figure 2). This occurs, despite the fact that all relationships between evidence and verdict that are displayed in the decisions of the

participants are 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, presumably dominant 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 does 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, 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 in the same direction but 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. The study shows that, although they understand the basics, many students and even the best law students in particular lack the skills to reach correct verdicts, when it comes to hard cases. It seems safe to conclude that they need a lot of training to handle uncertainty correctly. We find that many participants make decisions in an intuitive way. We agree with Guthrie et al. (2007) that rational deliberation needs to take a more prominent place in court rooms. While this is not sufficient to deal with uncertainty correctly, it would be an important step.

The reliance of many participants on intuition rather than the evidence-belief-decision sequence is intriguing. It would be interesting to examine how intuition works by modeling the relationships between decisions and available evidence. Because of the heterogeneity in behavior, this can only be done per individual. Because of the binary choice, much more data per individual are needed than we gathered here to perform such analyses. This is a direction for future research.

Literature

Bornstein, B. (2004). The impact of different types of expert scientific testimony on mock jurors' liability verdicts, *Psychology, Crime and Law* 10, 429-446.

Brink, A, L. Schomaker and M. Bulacu (2007). Towards explainable writer verification and identification using vantage writers. *ICDAR*, 824-828.

Buchanan, M. (2007). Conviction by numbers. Nature, 445, 254-255.

Clark, S.E. and G.L. Wells (forthcoming). On the diagnosticity of multiple-witness identification. *Law and Human Behavior.* Online First.

Dawd, Ph. (2005). Statistics on trial, *Significance* 2, 6-8.

Donnelly, P. (2005). Appealing statistics, *Significance* 2, 46-48.

Eddy, D.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* 249, 253-254.

Edwards, W. (1968). Conservatism in human information processing. In: B, Kleinmuntz (ed.), *Formal representation of human judgment.* Wiley.

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

Guthrie, C., J.J. Rachlinski and A.J. Wistrich (2001). Inside the judicial mind, *Cornell Law Review* 93, 1-43.

Guthrie, C. , J.J. Rachlinski and A.J. Wistrich (2007). Blinking on the bench: how judges decide cases, *Cornell Law Review* 86, 777-830.

Hartendorp, R.C. (2008). *Praktisch gesproken: Alledaagse civiele rechtspleging als praktische oordeelsvorming* (Practically spoken: Everyday civil procedure as practical decisioning, PhD thesis, Erasmus Universiteit Rotterdam.

Holt, C. and S.K. Laury (2002). Risk aversion and incentive effects. *American Economic Review* 92, 1644-1655.

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

Kahneman, D. and A. Tversky, 1972, Subjective probability: A judgment of representativeness, *Cognitive Psychology* 3, 430-454.

Kassin, S. M., R. A. Leo, C.A. Meissner, K.D. Richman, L.H. Colwell, A-M. Leach and D. LaFon (2007). Police interviewing and interrogation: A self-report survey of police practices and beliefs. *Law & Human Behavior*, *31*, 381-400.

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

Koppen, P.J. van. (2008). Blundering justice: The Schiedam Park Murder. In R.N. Kocsis, *Serial murder and the psychology of violent crimes* (pp. 207-228). Humana.

Pratt, J.W., H. Raiffa and R. Schlaifer (1995). *Statistical decision theory*. MIT Press.

McAllister, H.A. And N.J. Bregman (1986). Juror underutilization of eyewitness nonidentifications: theoretical and practical implications. *Journal of Applied Psychology* 71, 168-170.

Meester, R., M. Collins, R. Gill and M. Van Lambalgen (2006). On the (ab)use of statistics in the legal case against the nurse Lucia de B., *Law, Probability and Risk*, 5, 233-250.

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

Offerman, T., J. Sonnemans, G. van de Kuilen and P.P. Wakker (2008). A Truth-Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes. **???**

Saks, M.J. and R. F. Kid (1980). Human information processing and adjudication: trial by heuristics, *Law and Society Review* 15, 123-160.

Sonnemans, J. (1998). Strategies of Search. *Journal of Economic Behavior and Organization* 35, 309-32.

Sonnemans, J. (2000). Decisions and Strategies in a Sequential Search Experiment. *Journal of Economic Psychology* 21, 91-102.

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

Thompson, W.C. and E.L. Schumann (1987). Interpretation of statistical evidence in criminal trials, *Law and Human Behavior* 11, 167-187.

Tversky, A. and D. Kahneman (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5, 297-323.

van der Heide, W., F. van Tulder and C. Wiebrens (2007). Strafrechter en strafketen: de gang van de zaken, 1995-2006, Rechtstreeks 3, 7-72.

Wagenaar, W.A. (2006). *Vincent plast op de grond; nachtmerries in het Nederlandse recht*. Bert Bakker.

Way, V. 2003, Judicial fact-finding by judges alone in serious criminal cases, *Melbourne University Law Review*, 27, 2, 423-457.

A shortened version of the experiment can be played at
http:www.creedexperiment.nl/recht2/begin.html
in English or Dutch

The appendix with the instructions and
additional analyses can be found at
http://www.fee.uva.nl/creed/pdffiles/judicialfactfindingappendix.pdf