Judicial error by groups and individuals[◦]

Frans van Dijk
Netherlands Council for the judiciary

Joep Sonnemans
CREED, University of Amsterdam and Tinbergen Institute

Eddy Bauw
Netherlands Council for the judiciary and University of Amsterdam

March 2012

Abstract

In criminal cases judges evaluate and combine probabilistic evidence to reach verdicts. Unavoidably, errors are made, resulting in unwarranted conviction or acquittal of defendants. This paper addresses the questions (1) whether hearing cases by teams of three persons leads to less error than hearing cases alone; (2) whether deliberation leads to better decisions than mechanical aggregation of individual opinions; and (3) whether participating in deliberations improves future individual decisions. We find that having more than one judge consider cases reduces error effectively. This does not mean that it is necessary to deliberate about all cases. In simple cases many errors can be avoided by mechanical aggregation of independent opinions, and deliberation has no added value. In difficult cases discussion leads to less error. The advantage of deliberation goes beyond the case at hand: although we provide no feedback about the quality of verdicts, it improves individual decisions in subsequent cases.

◦ We thank the participants of the "Judgement day" at the Netherlands Council of the judiciary and Roel van Veldhuizen for their comments on an earlier draft of this paper.

1. Introduction

In this paper we focus on binary judicial decisions (acquit or convict the defendant in a criminal case) based upon multiple probabilistic evidence. Combining pieces of probabilistic evidence asks for sophisticated reasoning which is hard for many decision-makers. In courts, these decisions are sometimes made by an individual and sometimes by small groups like a panel of three judges. We compare experimentally the quality of decisions made by small groups after (or without) deliberation and individual decisions[1].

Judicial decision making, as a special case of decision making under uncertainty, offers specific challenges. The experiment reported by Sonnemans and Van Dijk (2011) shows high error rates for the decision problems judges face in criminal cases. Processing evidence requires assessing and combining of probabilities, which leads to many mistakes. Most are of the most serious kind: convicting innocent defendants. The shortcomings of human reasoning are becoming a more explicit problem in this field, as criminal evidence gets more technical and probabilistic due to the progress made by the forensic sciences.

Sonnemans and Van Dijk examined these issues in the context of individual decision making. As a result, their findings may offer a picture of flawed decisions that is too bleak. Charness et al (2007, p147) claim "that the experimental evidence indicating systematic deviations from the courses of action prescribed by normative models of decision making under risk and uncertainty, such as expected utility theory, are due, in part, to the artificial isolation imposed by the experimental setting. These violations tend to be less pronounced when social interaction is allowed". They find that errors are significantly less frequent in decisions by groups of three than in individual decisions for the case of Bayesian updating of information about lotteries. As the experiment of Sonnemans and Van Dijk dealt also with Bayesian updating, albeit in a much more complicated setting, this result is directly relevant. Furthermore, one may assume that not without reason in most countries judicial decisions especially in more difficult cases are not taken by a single judge, but by

[1] There are three reasons why we use a judicial (and not business frame) in our experiment. Two of the three authors (FvD and EB) work in the field of judicial decisions and this research was provoked by a (cost reducing) plan of the Dutch government to reduce the number of criminal cases that are decided by panels of three judges instead of a single judge. Furthermore, judicial organization is based on an explicit view of the merits of individual versus group decision-making. Finally, an (experimental) advantage of a judicial frame is that subejcts appear to understand the basic concepts of the decision problem quite well.

teams of judges, lay persons or combinations of both. For instance, in the Netherlands a panel of three judges hears more serious criminal cases, while the other criminal cases are adjudicated by a single judge. Other countries rely on larger teams of judges and lay persons (e.g. Germany) or on juries (e.g. US and UK) for trying serious crime. Lesser crimes are adjudicated by a single judge like in the Netherlands or by small teams of judges (e.g. three magistrates in the UK). In most countries appeals are heard by panels of judges. Thus, in legal tradition the belief that group decisions are better than individual decisions is deeply ingrained and formalized. Thereby it offers an important, practical case of individual versus group decision making.

In this paper we examine the impact of having decisions made by small groups instead of individuals on judicial error by extending the experiment of Sonnemans and Van Dijk. The three main questions we try to answer are: firstly, do groups perform better than individuals? Secondly, to what extent (if at all) does *deliberation* in groups reduces error rates? For this purpose we compare group decisions with mechanical aggregation of individual decisions of group members. Thirdly, we will look at effects beyond the current decision: does participation in group decision-making increase the quality of individual decisions in *subsequent* cases? As noted, the experimental task studied is judicial, but aspects of this task are very common in economic contexts where uncertainty is endemic, and different pieces of information have to be evaluated and combined, before decisions can be made.

In section 2 we briefly review current opinion among the legal profession in general and judges in particular about the merits of hearing cases alone and in panels, as well as the scientific evidence. Section 3 gives the hypotheses, while in 4 we set out our methodology and the design of the experiment. Section 5 gives the results. In section 6 we pay specific attention to the efficiency of decision-making. Section 7 concludes.

2. Literature

Views of the legal profession

Within the legal profession, the opinion is prevalent that teams of judges make better decisions than single judges, and clients share that view. This is documented in extenso by Baas et al. (2010) for the Netherlands. They conducted an extensive review of the legal literature and policy documents in the Netherlands, as well as a questionnaire among Dutch judges, and conclude that the dominant view is that the

probability of a correct decision increases when more judges hear a case (Baas et al., p41). Among the profession, it is also felt that not only the larger number of judges, but also the professional deliberation contributes to better decisions (Baas et al., p42). Considerations reported are: more aspects of the case are considered, smaller risk of (technical) error, reduced risk of a one-sided approach, better consideration of case law and current legal opinion and more awareness of society. Some of these considerations are tautological, but they elucidate that, while several reasons exist for having cases considered by more than one judge, reducing error is an important one. Judges also note that in simple cases with clear-cut evidence teams of judges add little value (Baas et al., p135). Related to our third research question about indirect learning effects of team decision making, judges argue specifically that participation in panels allows new judges to build up experience and learn from other judges, while participation is also relevant for experienced judges to expose them to peer review (Baas et al., p128).

A study for the UK compares the performance of lay magistrates and professional judges (Ipsos Mori, 2011). In roughly the same cases, the judges sit alone, while the magistrates hear cases in panels of three. The perceptions of clients and their lawyers, as well as judges, magistrates and staff were solicited. Sitting in a panel was seen as an important strength by many respondents. One reason is "fairness": hearing cases in teams results in more balanced decisions, as individual prejudice is corrected. But respondents feel also that teams reduce the potential for misunderstanding key points of a case. In the words of a defendant, quoted in the report:

> "When you get one on his own he could get the wrong end of the stick or misunderstand something. If there's a couple more then they can guide him; they'll see a different point". (Ipsos Mori 2011, p. 19)

In Baas et al. (2010) disadvantages of teams are also recognized by respondents: decision-making takes longer and costs more. The view is also expressed that leaving the case in the hands of one judge leads to greater responsibility and greater conscientiousness, and that could result in better decisions (Baas et al., p44). In the UK study, respondents state that judges work faster than panels of magistrates, because they sit alone, but also - and that is a confounding factor -

because of better legal knowledge. For the Netherlands as well as for the UK it is found that teams actually take much more time than judges sitting alone, but in both countries the comparison is flawed. In the Netherlands the cases differ qualitatively, while the judges are the same; in the UK the cases are the same, but the judges (or magistrates) differ.

Baas et al. conclude that no empirical basis exists for or against the claims about the quality of decisions, and call for more research. In the UK study quality of decision-making was considered to be highly subjective, and was therefore not empirically examined. Consequently, there is good reason to examine empirically whether judicial decision-making improves when decisions are not taken by a single judge, but by a panel of judges. This can contribute to a better understanding of (judicial) decision-making in general. In laboratory experiments the quality of decisions can be assessed unambiguously and the same cases can be presented to different individual or group decision makers, avoiding the major methodological problems of field studies.

Judges and juries

In the research about judicial decision making much attention has been given to the relative performance of judges and juries. One would expect that the differences between individual and group decision-making would have received much attention as well. However, in a review article Robbennolt (2005) notes that most of the comparative studies examine decision-making of individual judges and individual jurors. For the type of decision problem we study here, it is of interest that such studies find that judges as well as juries have difficulty understanding scientific and statistical evidence, underlining the relevance of the treatment of uncertainty in judicial decision making. As to the comparison of individual and group decisions Robbennolt gives centre stage to deliberation, but concludes that no general conclusion can be drawn about the impact of deliberation in juries on decisions. She notes that deliberation can lessen biases in some instances, and for instance leads to more complex reasoning about the evidence and arguments presented. It also reduces variability in decisions. However, group discussion may also worsen biases under certain conditions, and can result in more extreme judgments. She concludes, as above, that additional research is clearly warranted.

Social psychology and experimental economics

Social psychology has addressed extensively the question whether groups are more or less subject to biases than individual decision makers. According to Kerr et al. (1996), the general consensus of review studies is that, on average, groups outperform individuals in various decision tasks that have an objective correct solution. In their review Kerr et al. focus on the gray area between these, so called, intellective tasks and pure decision-making. Reviewing a host of studies, they find that there is no general answer to the question whether groups are more or less biased than individuals. Outcomes depend in particular on group size, magnitude of individual bias, type of bias (categories used are "sins of imprecision", "sins of commission" and "sins of omission"), and group process. The outcome is a fragmented picture of a very large number of biases humans are subject to, the effects of which depend on many factors. Kerr et al. note in particular the importance of the method groups use to reach decisions such as simple majority or consensus. In practice, simple majority seems to be used most often, and it is actually used by panels of judges in the Netherlands, when unanimity cannot be reached. It is one of the more simple mechanisms to analyze the transmission of individual bias to group bias, as it serves to make the more popular individual choices even more popular in groups (Davis, 1973).

Stepping back from the study of biases and focusing on the quality of decisions, the relative performance of groups has received much attention in experimental economics, starting with Cooper and Kagel (2005) which found that groups outperform individuals in signaling games. The experiment of Charness et al. (2007) is of particular interest to the present issues. In this experiment participants chose between risky prospects in lotteries, individually and in small groups (2 and 3 persons). Treatments differ in the potential presence of affect, the need for Bayesian updating and type of lottery. The experiment tests whether participants adhere to basic principles concerning the monotonicity of first-order stochastic dominance and Bayesian updating. The main results are that both principles are violated by a substantial number of participants, deciding individually, and that the number of violations is smaller in groups, especially, of three persons. Blinder and Morgan (2005) also find that group decisions are on average better than individual choices for a purely statistical decision problem as well as a similar problem but framed in the context of monetary policy. Surprisingly, they also find that group decisions do not

take more time than individual decisions, and that decisions requiring unanimity do not take more time than decisions under majority rule.

Judicial decision-making in teams is generally based on deliberation. As described above, judges believe that deliberation as such adds to the quality of decisions over and above the mere aggregation of opinion. This is, however, not self-evident because many psychological studies have shown that (intensive) communication within groups can lead to more extreme positions (group polarization) which may lead to bad decisions. Surowiecki argues in his bestselling book "The wisdom of crowds" (2004, chapter 3) that groups perform better only if the members form their opinions independently. Lorenz et al (2011) shows that social influence may lead to convergence of estimates and therefore reduces diversity, even when minimal interaction is allowed, while confidence in the estimates increases. The authors suggest that this threatens the wisdom of crowds, resulting from the mechanical aggregation of independent opinions, although they do not find an increase of collective error as such. On the other hand, several studies show that group decisions are better than individual decisions as a result of discussion within groups (Casari, Zhang and Jackson, 2010; Cooper and Sutter, 2011). Kocher and Sutter (2005) finds that groups learn faster than individuals in a guessing game. This does not prove, however, that discussion improves decisions when compared with mechanical aggregation of opinion. Lombardelli et al (2005) experimentally studies monetary policy decisions by individuals and groups and find that although groups perform better than individuals, there is no difference in performance between groups that voted with or without discussion. This suggests that it is not a foregone conclusion that allowing deliberation leads to better decisions than mechanical aggregation of individual opinions.

There is some evidence that participating in a group decision task improves subsequent individual decision making in similar tasks (our third research question)[2]. Using mathematical problems, Stasson et al (1991) find that individuals perform better after having previous taken part in a group problem-solving task. However, it is not clear from the description of the experimental procedures whether the participants learned the accuracy of the group solution before making the individual decisions.

---

2 This is called "general group to individual transfer" in psychological jargon; see Laughlin (2011, chapter 7).

Laughlin et al (2008) use letters to numbers problems[3] in which by construction the correctness of answers is provided. This is different from real life situations (and also the design of our experiment) where direct feedback about the correctness of a decision is rare[4]. The only study we could find in which no feedback was provided in the group decision stage of the experiment is study 1 in Maciejovsky and Budescu (2007). They use the Watson test (logical reasoning) and find that individuals learn from the group deliberation.

3. Research questions and hypotheses

As in Sonnemans and Van Dijk (2011), we compare experimentally the decisions of actual decision makers with risk neutral optimal decisions, which we denote as the normative model. Here, the key issue is the comparison of group and individual decision-making, and the explanation of any differences found. The normative model will be set out in the following sections, but we can now already formulate the hypotheses to be answered by the experiment.

Hypothesis 1: Groups perform better than individuals.
Adjudicating cases by a team of judges leads to less error than hearing cases by a single judge, where error is defined as any deviation from the outcome of the normative model. We expect this effect only in complicated, and not in simple cases.

Hypothesis 2: Both aggregation and deliberation play a role in the quality of group decision making.
Decisions of small groups are more accurate than individual decisions as a result of – without any deliberation or other interaction – aggregation of individual decisions (wisdom of (small) crowds). In addition to this effect, we expect that decisions by groups are also more accurate than individual decisions as a result of deliberation.

[3] In this problem the number 0-9 are randomly coded to the letters A-J. The objective is to indentify the mapping in as few trials as possible (like the Mastermind game) On each trial the problem solver proposes an equation in letters (e.g., A + D=?) and receives the answer in letters (e.g., A + D=B), proposes one specific mapping (e.g., A=3), receives the answer (e.g., True, A=3), or proposes the full mapping of the 10 letters to the 10 numbers.

[4] For example, a board of directors who decide whether to merge with another firm will never know what the outcome would have been if the other decision was made. The same goes for judges who rarely get unequivocal confirmation or negation that a decision was right.

Hypothesis 3: individual decisions improve after participating in group decision-making

Individuals learn from taking part in groups, and as result their decisions when they individually hear cases become more accurate.

4. Methodology and design

We use a decision problem that has a fully determined solution, given the incentives of the decision makers: the decision to condemn or acquit a defendant on the basis of given probabilistic evidence which to that end needs to evaluated and combined. The incentives are controlled by providing financial pay-offs. Each possible outcome of decisions earns participants a pay-off. This means that for given risk attitude optimal decisions exist. The risk neutral optimal decision is denoted as the outcome of the normative model. We will first discuss the decision problem (4.1), and after that the methodology (4.2) and procedures (4.3)

4.1 Decision problem

The decision problem concerns the adjudication of criminal cases in their most elementary form. The defendant is guilty or not guilty, and all evidence (incriminating or exonerating) is directly informative. For details and discussion we refer to Sonnemans and Van Dijk (2011).

Errors and incentives

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

- Convict an innocent defendant, which is a grave injustice to the individual concerned and leaves the real perpetrator at large at the risk of repetition.
- 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.

Table 1 gives the incentive structure.

		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

Impartiality implies that *a* and *d* should be equal: the judge should not have a preference for one of these outcomes. This is not the case for *b* and *c*, where it would seem that $b << c$ (in both cases the real perpetrator is still at large, but a wrongful conviction has high additional costs for the innocent person convicted). The weights judges actually attach to these outcomes are fundamentally implicit to their functioning and cannot be known with any precision. Therefore, we impose them in the experiment. As we are only interested in the comparison of actual and optimal decisions, this does not limit the generality of the conclusions.

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 $ap+b(1-p)=cp+d(1-p)$ with p the probability of guilt. Or: $p = (d-b)/(a-b-c+d)$. In the experiment a=d=100, b=-1500 and c=-300 euro cents and thus the risk neutral decision maker should convict only when the probability of guilt is larger than 80%.[5]

Evidence and uncertainty

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

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

[5] 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.

(1)

In words: Posterior odds equals Prior odds multiplied by the Strength of evidence.

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). As convictions cannot be based on a single piece of evidence in most legal systems, generally the probabilities associated with different pieces of evidence have to be combined. The reality is that in some cases evidence will be contradictory. To allow for separate pieces of evidence, Equation (1) can be generalized, denoting the strength of a piece of evidence i as E_i, and assuming independent evidence:

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

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

Table 2 provides the structure of the evidence, as 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, of the same or other type(s).

The procedure to generate cases and associated evidence was as follows. First, whether the defendant was guilty or not was randomly determined with equal probability of innocence and guilt. 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 2. 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 (see Appendix 1 for screen shots).

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 2. Strength of evidence. In the experiment incriminating evidence was printed in red and exonerating evidence in blue and the size of the codes differed with the strength (see Appendix 1).

4.2 Methodology

In Sonnemans and Van Dijk (2011) this decision problem was used by having participants decide 30 cases individually. In the experiment reported here the same 30 cases were presented to three-person groups, the frequently used size of judicial panels, in the same sequence. Participants were randomly assigned to these groups of three persons. In each of the 30 cases participants first had to decide individually to convict or acquit the defendant and to assess the probability of guilt. Having done that, they had to decide the cases in their groups. To that end they were asked to discuss the cases in chat sessions. Each group member had to make at least one contribution to the group discussion. Having made a contribution, each group member could quit the chat session at his chosen moment and propose to convict or acquit the defendant, giving also his assessment of the probability of guilt. The group decision was then determined by simple majority. The group assessment of the probability of guilt was calculated as the median of the individual probability assessments. This has the advantage that participants do not have an incentive to give assessments that diverge from their actual beliefs for strategic reasons. For each case one of the four decisions (individual and group decisions to convict or acquit and the individual and group assessment of probability of guilt) was randomly selected and paid out at the end of the experiment. The probability of guilt was incentivized using a quadratic scoring rule (payoff between 0 and 1 euro).

Table 3 distinguishes the five different types of decisions we need to test the hypotheses of the previous section. Two of these decisions (IndGroup and GroupGroup) are directly observed in this experiment, the individual decision in the individual treatment (IndInd) is observed in a previously reported individual experiment (Sonnemans and van Dijk, 2011), and two decisions are constructed based upon the individual decisions in the individual and group treatments (ConGroupGroup and ConGroupInd). These constructed group decisions are determined by applying simple majority rule to the individual decisions. In the group treatment, this would have been the group decision, if we had not allowed deliberation to take place. In this treatment we use the existing group formation (which enables within subjects statistical tests), in the individual treatment we assign subjects randomly to three-person groups.

Code	*Description*	*Experiment*
IndGroup	Individual decision in group treatment	This study
GroupGroup	Group decision after discussion	This study
ConGroupGroup	Constructed group decision in group treatment based upon IndGroup	This study
IndInd	Individual decision in individual treatment	The individual treatment is reported in: Sonnemans & van Dijk 2011
ConGroupInd	Constructed group decision in individual treatment based upon IndInd	The individual treatment is reported in: Sonnemans & van Dijk 2011

Table 3. Treatment variables

The quality of decisions by individuals and groups (hypothesis 1) are studied by comparing GroupGroup with IndGroup and IndInd. The difference in quality between individual and group decisions can be disentangled (hypothesis 2) by distinguishing the effect of the wisdom of crowds which is the difference between IndGroup and ConGroupGroup, and the result of deliberation. The effect of deliberation can be measured in two ways. If we focus on a case-by-case difference, the right comparison is between the decision without deliberation in the group treatment (ConGroupGroup) and the same decision after deliberation (GroupGroup). Alternatively, we can

compare the wisdom of crowds of participants who *never* deliberate (ConGroupInd) with the decisions after deliberation (GroupGroup). Any difference in outcome between these two comparisons can be attributed to learning effects in the group treatment, which brings us to the third hypothesis. While the actual guilt or innocence of defendants was not revealed during the experiment and could not be a source of learning, participants could learn from each others' arguments in previous cases (hypothesis 3). To study the learning effect of the group discussion we compare the individual decisions IndGroup and IndInd.

4.3. Procedures

Computer screens and the instructions are given in Appendix 1 and the reader can anonymously participate in an online, individual version of the experiment at www.creedexperiment.nl/recht2/begin.html.

Participants had to decide 30 cases, with which they could earn money. In addition to the earnings to be discussed below, all participants earned a salary of 100 euro cents per case. 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 are given in Appendix 2.

To guarantee their understanding of the experiment, participants had to answer computerized questions individually 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 again individually, 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 were only given at the end of the experiment.

For every case, participants reported the subjective probability that the accused was guilty, and made the decision to convict or acquit. The decision was rewarded according to table 1 with, as mentioned before, in euro cents, 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., 2009). This procedure prevents hedging behavior by participants. All participants received the same cases and evidence. The risk neutral decision maker should only convict the accused when the evidence points to a

probability of guilt higher than 80%. This occurred in 8 of the 30 cases. In section 4.1, we specified already the way in which individual and group decisions were made.

Participants The experiment took place at the CREED laboratory of the University of Amsterdam. Participants were social sciences students, mainly in economics and psychology. In total 99 students participated in the experiment, earning on average 32 euro in about two hours. In the experiment of Sonnemans and Van Dijk (2011), with which results are compared, also law students and candidate judges participated. For the comparison only the data for participants studying social sciences (122 students) are used.

5. Results

We will follow the hypotheses formulated in section 3. All statistical tests are non parametric (Mann-Whitney) and the level of observation is the average over the cases of individuals or groups.

5.1 The quality of group versus individual decision-making

Figure 1 shows the relationship between the proportion of convictions and the objective probability of guilt. Recall that if the probability is smaller than 80% rational risk neutral decision makers will acquit the defendant. Else, they will convict the defendant. In the figure the group decisions following deliberation (GroupGroup), individual decisions of participants in the group condition (IndGroup) and individual decisions in the individual condition (IndInd) are depicted. The group decisions (green) are closer to the optimal decisions (black broken line) than the individual decisions (blue and red). Also, individual decisions in the group condition (blue) are better than the individual decisions (red), indicating that group discussions improve subsequent individual decisions. Table 4 gives the corresponding average number of errors.

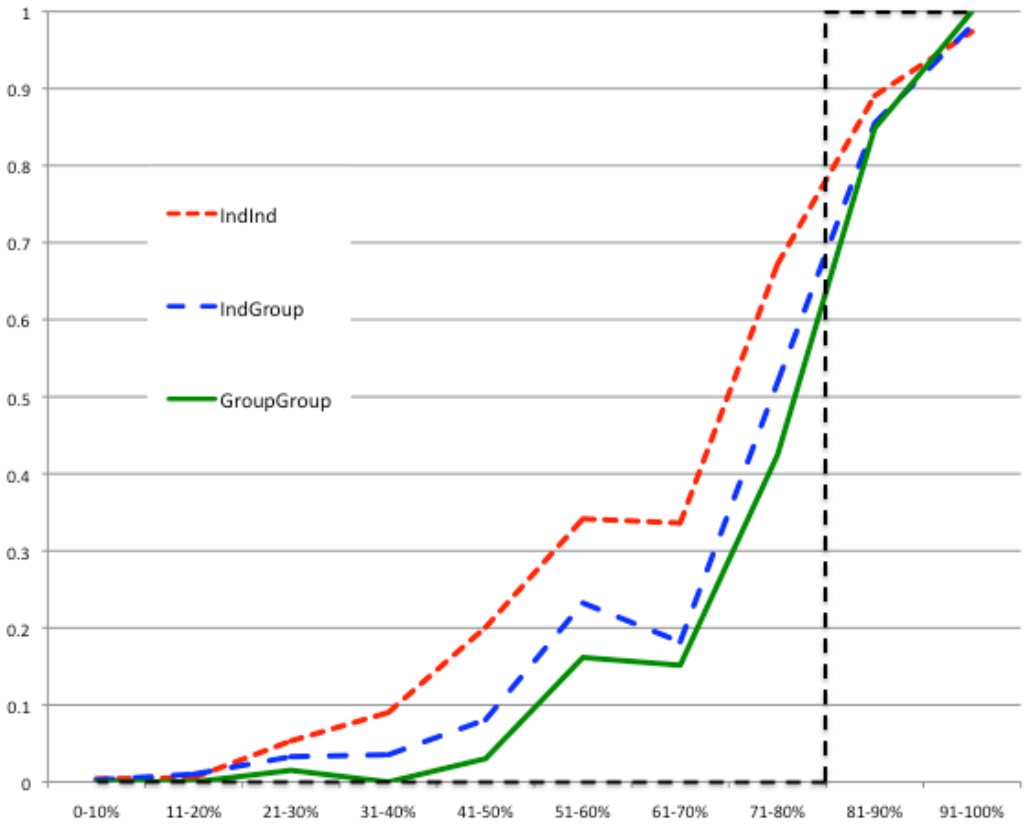


Figure 1. Proportion of convictions (vertical axis) and the objective probability of guilt. The black broken line gives the optimal decisions.

	# wrong Acquittals	# wrong Convictions	Total errors
GroupGroup	0.45	2.03	2.48
IndGroup	0.53	2.84	3.36
IndInd	0.46	4.20	4.66
Tests p-values			
GroupGroup vs. IndGroup (Wilcoxon)	.407	**.000**	**.000**
GroupGroup vs, IndInd (MW)	.634	**.000**	**.000**
IndGroup vs. IndInd (MW)	.429	**.001**	**.001**

Table 4. Average number of errors of both types and in total, in 30 cases for all participants, and statistical tests. A decision is a wrong acquittal (conviction) when the accused is acquitted (convicted) while the objective probability of guilt is larger (smaller) than 80%. p-values smaller than 0.05 are printed in bold.

Hypothesis	Treatment	*Objective probability of guilt* *0-40%* (13 cases)	*40-80%* (9 cases)	*80-100%* (8 cases)	All cases
	IndInd	33.5	128.4	10.1	55.7
	ConGroupInd	5.3	108.4	1.6	35.2
	IndGroup	20.2	81.8	10.3	36.0
	ConGroupGroup	2.6	68.3	4.5	22.8
	GroupGroup	5.4	56.7	7.1	21.2
	Tests p-values				
H1	IndInd vs. GroupGroup (MW test)	**0.017**	**0.000**	0.827	**0.000**
H1	IndGroup vs GroupGroup (Wilcoxon test)	**0.003**	**0.000**	**0.050**	**0.000**
H2	IndGroup vs ConGroupGroup (Wilcoxon test)	**0.008**	**0.008**	**0.000**	**0.000**
H2	IndInd vs ConGroupInd (Wilcoxon test)	**0.000**	**0.000**	**0.000**	**0.000**
H2	ConGroupGroup vs GroupGroup (Wilcoxon test)	0.414	**0.018**	**0.024**	0.231
H2	ConGroupInd vs GroupGroup (MW test)	0.859	**0.004**	**0.023**	**0.021**
H3	IndInd vs IndGroup (MW test)	0.057	**0.003**	0.767	**0.002**
-	ConGroupInd vs ConGroupGroup (MW test)	0.423	**0.024**	0.409	0.116
-	IndInd vs ConGroupGroup (MW test)	**0.000**	**0.000**	**0.000**	**0.000**
-	ConGroupInd vs IndGroup (MW test)	**0.009**	0.425	**0.000**	0.641

Table 5: Average error per case, defined as the difference of expected earnings of actual decisions and expected earnings of optimal decisions, in cents, and statistical tests. p-values smaller than 0.05 are printed in bold. The last three rows are not directly related to the hypotheses but displayed for completeness.

As incentives are of a financial nature, error can best be expressed in costs per case, defined as the difference between the expected earnings of actual decisions and the expected earnings of optimal decisions, see table 5. For all 30 cases the costs per case are 21.2 cents for group decisions, 36.0 cents for individual decisions in the group condition and 55.7 cents for the individual decisions in the individual condition. The differences are highly significant. To compare within subjects individual and group decisions in the group condition the Wilcoxon test is applied (p=0.000).

Errors can occur due to a wrong assessment of the probability of guilt and to a wrong decision to convict or acquit a defendant for a given assessment of the probability of guilt. In Sonnemans and Van Dijk (2011) it was found that the main source of error are the decisions, as the assessment of probability is reasonably accurate. We find this in the group condition as well. The average error in cents in the assessment of the probability is 1.9 cents in the group decisions, and 3.1 and 2.8 cents for the individuals in the group and the individual treatment. The groups do statistically significant better than the individuals (both p's <0.01), but the difference

between the two types of individual decisions is insignificant (p=.68. Obviously, the main cause of error lies in the decisions, and we will not discuss the probability assessment any further.

Aggregating cases in three ranges of objective probability of guilt highlights the differences for cases of different difficulty (Figure 2). For small as well as large probability of guilt decisions are simple, and the differences between the conditions are small. In the middle range decisions are complicated, and there the differences are large. In this area the optimal decision is to acquit defendants, and, consequently, all errors are of the worst kind.

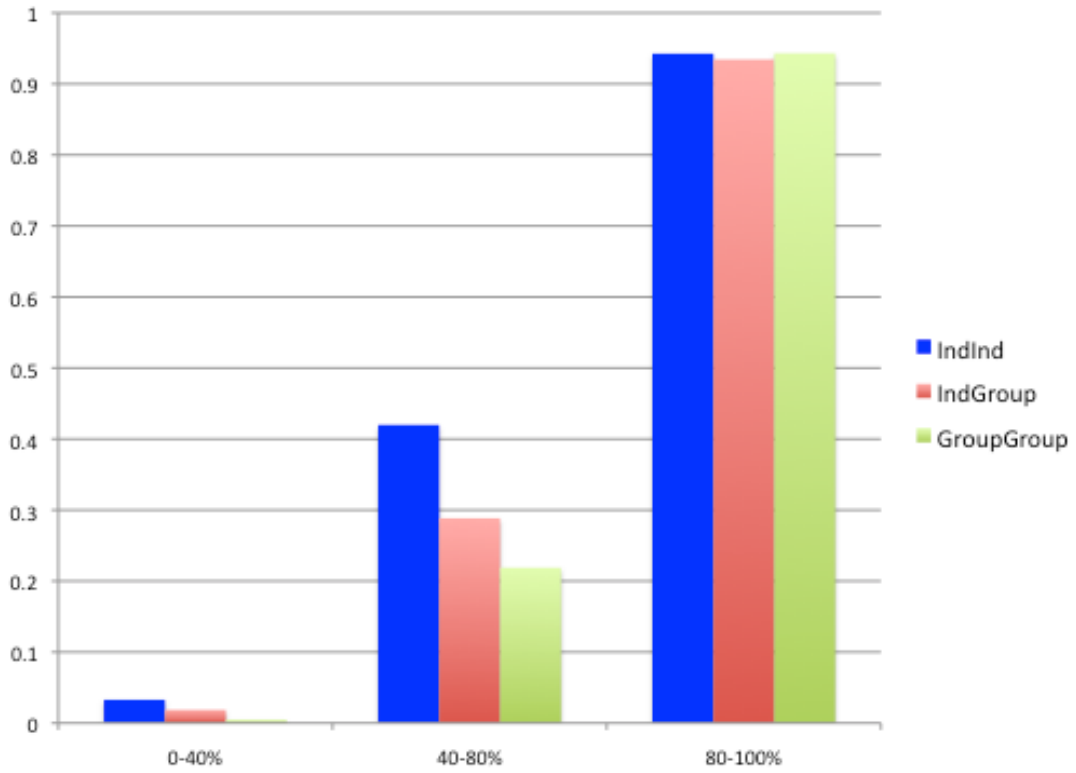


Figure 2. Proportion of convictions for three categories of objective probability of guilt.

By having a team decide instead of one judge, error cost is reduced most for the difficult cases (56.7 vs. 128.4), but, the reduction is also large for the relatively easy cases with low probability of guilt (5.4 vs. 33.5). In this range error is also of the worst kind (convicting an innocent defendant). Avoiding such rare but very costly errors has a large return. Consequently, reducing error in simple cases by hearing these by a panel of judges has a substantial return by avoiding unwarranted

convictions. For simple cases with high probability of guilt individual decisions have little error (10.1), and there is hardly any room for improvement (7.1). This difference is statistically insignificant.

We can conclude that in line with hypothesis 1 groups perform better than individuals. Against our expectation, this effect is not limited to the difficult cases (objective probability of guilt 40-80%): also for a range of simple cases groups perform better.

5.2 The effect of aggregation of opinion and deliberation

There are two ways to compare the accuracy of decisions by groups based on – without any deliberation or other interaction – aggregation of individual decisions (wisdom of (small) crowds) and individual decisions. First, in the group condition individual decisions (IndGroup) can be compared with the constructed decisions derived by aggregating the individual decisions using simple majority (ConGroupGroup). Second, the individual decisions in the individual condition (IndInd) can be compared with constructed group decisions derived by aggregating individual decisions in the same way (ConGroupInd). In the group condition the groups exist; in the individual condition the groups are formed randomly.

Table 5 gives the numerical values for the relevant decisions, and also provides the statistical tests. The differences between individual decisions and the same individual decisions, aggregated in groups of three participants, are highly significant. In the individual condition average error costs are 55.7 cents for individual decisions (IndInd) and 35.2 cents for constructed group decisions (ConGroupInd). In the group condition average error costs are 36.0 cents for individual decisions (IndGroup) and 22.8 cents for group decisions (without deliberation, ConGroupGroup) for all cases. In all three ranges of objective probability of guilt, the differences are large. Decisions improve substantially by mechanically combining the individual decisions. In Table 5 also both types of mechanical group decisions are compared. Only for difficult cases the differences are significant; in the other ranges error is very small for both types of group decisions. The wisdom of crowds manifests itself strongly, and this part of the hypothesis stands.

Next, we study whether deliberation has an additional effect on the quality of decisions. Table 5 shows that for the difficult cases (objective probability of guilt ranging from 40 till 80%) error is significantly less in group decisions that are based

on deliberation (GroupGroup) than in mechanically generated group decisions (ConGroupGroup): 56.7 cents versus 68.3 cents. For the other cases this is not so; with respect to simple cases with high objective probability of guilt costs are even significantly higher with than without deliberation. Over all cases the differences are small and insignificant. It may be argued that for examining the effect of deliberation a better comparison is between mechanically generated group decisions of the decision-makers who have no experience of deliberation in previous periods (ConGroupInd) versus the decisions after deliberation in the group treatment (GroupGroup). This comparison also rules out the effect of previous deliberations. The differences are larger and also statistically significant if we consider all thirty cases. However, for the cases with a high probability of guilt we find a small but statistical significant difference in the other direction: ConGroupInd performs somewhat better than GroupGroup for these cases.

We take a closer, explorative, look at cases where groups (GroupGroup) make a different decision than the (constructed) wisdom of crowds groups (ConGroupGroup). Table 6 shows than in 955 (96%) cases the decision is the same. In 8 cases the groups convict while the majority of the individual decisions (and thus the constructed group decisions) are acquittals; in all 8 cases (from 7 different groups) the objective probability was lower than 80% and the right decision was to acquit; deliberation led to a deterioration of the quality of the decisions. In 27 cases the majority of the individual decisions was to convict, while the group decision was to acquit. In 23 of these 27 cases the right decision was to acquit, and the group made the correct decision. When the objective probability of guilt is smaller than 40%, in two (one) cases deliberation improves (worsens) the decision. In the more difficult cases with an objective probability between 40% and 80% deliberation improves (worsens) decisions in 20 (6) cases. Finally, there are 6 cases with an objective probability larger than 80% in which the group acquits while the majority of the individual decisions was (rightly) to convict. Apparently, deliberation helps in difficult cases, but does not improve decisions in simple cases, and may even lead to a deterioration of some decisions. Mechanical aggregation of individual decisions reduces error already to a very large degree. Deliberation can then easily lead astray.

We can also conclude from table 6 that there is a strong tendency towards unanimity (94% in GroupGroup versus 82% in ConGroupGroup). Deliberation causes

many participants to change their mind, whether by conviction, trust in other group members or due to group pressure.

			GroupGroup				
			acquit		convict		
			0	1	2	3	Total
ConGroupGroup	acquit	0	554	2	0	0	556
		1	75	16	3	5	99
	convict	2	14	9	23	37	83
		3	2	2	1	247	252
		Total	645	29	27	289	990

Table 6: Cross table of individual opinions in groups and group decisions for the two types of group decisions. In the columns are the group decisions (GroupGroup) and the number of opinions in favor of conviction; the rows show the constructed group decisions (ConGroupGroup) and the number of individual opinions to convict.

An analysis of the deliberations shows the ways in which errors are corrected. Firstly, many elementary mistakes are eliminated. Most of these concern the understanding of the evidence and the decisions to be made[6]. Secondly, most groups discuss the probability of guilt in each case. Group members generally tell the others their estimates. Evidently wrong estimates are corrected, and groups often settle on some average of the remaining individual estimates. Thirdly, groups discuss whether to acquit or convict the defendant. Reasoning with respect to the relationship between probability of guilt and verdict is mostly unsystematic. Only 8 groups explicitly discuss a minimum probability of guilt necessary to convict a defendant. This threshold varies between 65% and 85%. While a threshold is often not mentioned explicitly, a reluctance to convict defendants without (very) strong evidence shows in the chats. In many groups members keep repeating the high cost of convicting innocent defendants to each other, and discussions often end with the safe course to acquit the defendant. Note that this cautious behavior of groups will decrease errors in the difficult cases (where acquittal is the right decision) but will increase errors when

[6] This happens in 10 groups.

the objective probability of guilt is between 80 and 100% (and conviction is the right decision).

5.3 Individuals learn from taking part in panels

Hypothesis 3 can be tested by comparing IndInd (individual decisions in the individual condition) with IndGroup (individual decisions in the group condition). In the group condition participants can learn from the arguments of other group members, as mentioned before (feedback is only given at the end of the experiment in both cases). Table 5 shows large differences in error; the costs are 55.7 cents in IndInd and 36.0 cents in IndGroup. For the middle range of difficult cases and the range of simple cases with low probability of guilt costs are respectively 36% and 40% lower. The differences are significant except for the cases with high probability of guilt, where again there is little room for improvement. We can conclude that deliberation in groups helps to improve subsequent individual decisions.

6. Efficiency of decision-making

As discussed in section 2, a disadvantage of decisions by panels of judges is that they take longer and cost more than decisions by single judges. In our experiment the time needed for group decisions consists of the time needed to reach an individual decision plus the time needed to reach a group decision. The average time to reach an individual decision is 29.3 seconds in the group condition. The deliberation lasted for 60.4 seconds on average. In the individual condition decisions took 24 seconds for the here relevant group of social science students. To reach group decisions takes 4.5 times longer and costs 13.5 times more labor. It is likely that decision-making processes in general are more interactive from the start, and thus an individual phase may not exist to the extent it does in our experiment. However, judges cannot be expected to enter discussions ill prepared. Even disregarding the individual phase, decisions take longer. We, therefore, do not find the result of Blinder and Morgan (2005) that groups do not need more decision time than individuals. Whether the reduction of error that we established weighs up against these costs, cannot be answered within our experiment. However, the results raise some relevant points. The largest, absolute reduction of error is found for difficult cases. For difficult cases costs are reduced from an average of 128.4 cents in purely individual decisions to 56.7

cents in group decisions after deliberation. For simple cases with low probability of guilt costs are reduced from 33.5 cents tot 5.4 cents and for simple cases with high probability of guilt from 10.1 cents to 7.1 cents, a difference which is insignificant. This would point to assigning panels of judges only to difficult cases. However, we also found that simple cases with low probability of guilt benefit by being heard by teams of judges, because of the costs of unwarranted conviction. The relatively simple remedy of targeting misses these cases.

We also found that mere aggregation of independent, individual decisions leads to a large reduction of error, and that deliberation only further reduces error in difficult cases. For simple cases with low probability of guilt costs are nearly equal for group decisions based on independent, individual decisions (ConGroupInd) and group decisions (GroupGroup) after deliberation (5.3 and 5.4 cents), while for simple cases with high probability of guilt the aggregation of independent, individual decisions leads even to better results (1.6 versus 7.1 cents). This suggests a nuanced approach. Difficult cases should get the full treatment of deliberation, while simple cases could benefit from having several judges examine the cases, but without deliberation. As a result, individual decision time in these simple cases would not increase, and, consequently, total time spent would only increase three fold when cases are not heard by a single judge, but by three judges.

The question then is of course how to distinguish between difficult and simple cases. One option is to screen cases up front for complications in the evidence, which is feasible, but costs effort in itself. Another possibility which in practice is used in several countries is to attach large value to confessions as evidence: in case the defendant has confessed, the procedure is simplified. However, false confessions are not infrequent[7], and in those instances lead to very complicated cases. Another idea would be not to focus on the difficulty of cases, but on the consistency of individual decisions. If all three decisions are unanimous, deliberation could be skipped. This would mean in our experiment that only in 18.4% of cases deliberation would take place. In those cases deliberation would take 84 seconds on average, and time spent on the cases by the three judges together would be 5.6 multiplied by the time a single judge needs to decide the cases. Decision error would hardly increase (21.8 vs. 21.4 cents). While this does not promote critical evaluation of the evidence, it is efficient.

[7] False confessions recently contributed to a number of erroneous verdicts in major cases (Van Koppen, 2008 and Meester et al. 2006).

Another issue is that in the experiments groups do not operate in an efficient manner. Only 3 of the 33 groups determine a strategy on how to analyze and decide the cases at the beginning[8]. Five other groups[9] arrive much later at a more or less shared strategy. The other groups do not articulate a strategy, and discussions have a repetitive pattern. There is much room for improvement. It seems that training could improve decisions and, in particular, the efficiency with which these are reached.

7. Conclusions

We examined the adjudication of lawsuits as a specific instance of decisionmaking under uncertainty, in which deciding alone and in small groups both occur. It is also a situation in which the optimal deployment of the two methods is of great practical importance. The results are, however, also relevant for other decision situations. The experiment shows that adjudicating cases by groups of three persons leads to less error than hearing cases by a single person. Especially, error of the worst kind, convicting an innocent defendant, occurs less frequently. Hearing cases by groups improves decision making in complicated cases and also in simple cases with low probability of guilt. Much of this gain can be realized by having three persons evaluate a case independently, and then aggregating their individual decisions by simple majority. In this way many individual errors are filtered out. We found that deliberation further improves decisions in difficult cases, in which the probability of guilt is neither strongly incriminating or exonerating, and thus evidence and consequences of unwarranted acquittal or conviction need to be weighted carefully. This cautiousness of groups works out well in the difficult cases, but also leads to acquittals in those cases where a risk-neutral decision-maker woud convict. The overall effect is a large reduction of decision costs.

As the deliberations took place in the form of chats and can therefore be analyzed afterwards, we have observed how decisions are made and how errors are avoided. Elementary misunderstandings and mistakes are corrected in many groups. Furthermore, it is striking that only a few groups determine quantitatively a minimum probability of guilt needed for conviction and compare that minimum with an

[8] I.e. during the first three cases. The first case is very simple, and does not require analysis.
[9] The interpretation of what constitutes a strategy is arbitrary, but five groups develop a general method for deciding cases.

estimate of the probability of guilt, based on the evidence. The vast majority of groups simultaneously discuss the probability of guilt in view of the evidence and the verdict, during which they very often point out to each other the dire consequences of convicting an innocent defendant in terms of the costs for themselves. Although this type of reasoning is not sophisticated, many errors are avoided in difficult cases.

Opinions generally converge within groups, resulting in an increase of the number of unanimous decisions from 82% to 94%. Recall that many of the 30 cases are necessarily simple, and for that reason lead to equal individual decisions. This convergence could well be an advantage of discussing cases, as defendants are more likely to accept unanimous than divided verdicts, and judges are more confident about their decisions.

Groups make better decisions than individuals, and the effect is not limited to the decision at hand. Importantly, we find that subsequent individual decisions improve by taking part in group deliberations about (difficult) cases. Participation in deliberating panels has educational value.

Finally, we looked at the efficiency of decision-making. Group decisions take more time and are thus more costly than individual decisions. The time to reach a decision is longer, and of course not one but, in our experiment, three persons are involved. The question whether the reduction of error warrants the extra costs cannot be answered experimentally. Limiting the use of panels to difficult cases, assuming these can be identified beforehand, focuses effort on where it is most effective, but it misses many possibilities to reduce error in simple cases for which the probability of guilt is relatively low. We have shown, however, for those simple cases that time consuming group discussions do not reduce error, when compared with mere aggregation of independent, individual decisions. Consequently, an interesting option could be to restrict group deliberation to difficult cases, and in other cases have three judges independently form an opinion and without allowing deliberation aggregate their opinions mechanically by simple majority.

To conclude, group decisions out perform individual decisions. Within groups, deliberation has important positive effects: it leads to less error in difficult cases, and improves the quality of individual decisions in subsequent cases. However, deliberation is not useful in simple cases, and may even be counterproductive then. Our findings are largely consistent with the prevalent view of the legal profession,

and confirm the wisdom of legal tradition in allocating cases to individual judges and teams of judges.

Literature

Baas, R., L. de Groot-van Leeuwen and M. Laemers (2010). Rechtspraak: samen of alleen; over meervoudige en enkelvoudige rechtspraak. Research memoranda 6-5, Raad voor de rechtspraak.

Blinder, A. and J. Morgan (2005). Are two heads better than one? An experimental analysis of group versus individual decision making. *Journal of Money, Credit, and Banking* 37, 789-811.

Casari, M., Zhang, J. and Jackson, C. (2010). Do Groups Fall Prey to the Winner's Curse? IEW Working Paper 504. Institute for Empirical Research in Economics - University of Zurich.

Charness, G., E. Karni and D. Levin (2007). Individual and group decision making under risk: an experimental study of Bayesian updating and violations of first-order stochastic dominance. *Journal of Risk and Uncertainty* 35, 129-148.

Cooper, D. J. and Kagel, J. H. (2005). Are Two Heads Better Than One? Team versus Individual Play in Signaling Games. *American Economic Review*, 95(3): 477-509.

Cooper, D.J. and M. Sutter (2011). Role selection and team performance. IZA DP 5892. Institute for the Study of Labor.

Davis, J.H. (1973). Group decision and social interaction: a theory of social decision schemes. *Psychological Review*, 80, 97-125.

Ipsos Mori (2011). The strengths and skills of the judiciary in het magistrates' courts. Ministry of Justice Research Series 9/11, www.justice.gov.uk/publications/research.htm

Kerr, N.L., R.J. MacCoun and G.P. Kramer (1996). Bias in judgment: comparing individuals and groups. *Psychological Review* 1996, 103/4, 687-719.

Kocher, M. and M. Sutter (2005). The decision maker matters: individual versus team behavior in beauty-contest games. *Economic Journal*, 115, 200-223.

Laughlin, P.R. (2011) *Group Problem Solving*. Princeton: Princeton University Press.

Laughlin, P.R., H.R. Carey and N.L. Kerr (2008). Group-to-Individual Problem-Solving Transfer *Group Processes & Intergroup Relations* 11, 319-330

Lombardelli, C., J.Proudman and J. Talbot (2005). Committees versus individuals: an experimental analysis of monetary policy decision making. *International Journal of Central Banking* 1, 181-205.

Lorenz, J., H. Rauhut, F. Schweitzer and D. Helbing, (2011) How social influence can undermine the wisdom of crowd effect, *PNAS* 108, 9020-9025, www.pnas.org/cgi/doi/10.1073/pnas.1008636108.

Maciejovsky, B. and D.V. Budescu (2007). Collective induction without cooperation? Learning and knowledge transfer in cooperative groups and competitive auctions. *Journal of Personality and Social Psychology*, 92, 854-870.

Martin, A.W. and D.A. Schum (1987). Quantifying burdens of proof: a likelihood ratio approach, *Jurimetrics Journal* 27, 383-402.

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.

Offerman, T., J. Sonnemans, G. van de Kuilen and P.P. Wakker (2009) A Truth-Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes *Review of Economic Studies* 76, 1461-1489

Robbennolt, J. (2005). Evaluating juries by comparison to judges: a benchmark for judging? *Florida State University Law Review* 32, 469-509.

Sonnemans, J. and F. van Dijk (2011). Errors in judicial decisions: experimental results. *Journal of Law, Economics and Organization*, first published online, http://dx.doi.org/10.1093/jleo/ewq019.

Stasson, M.F., T. Kameda, C.D. Parks, S.K. Zimmerman and J.H. Davis (1991). Effects of assigned group consensus requirement on group problem solving and group members' learning. *Social Psychology Quarterly* 54, 25-35.

Surowiecki, J. (2004). *The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations.* New York: Doubleday.

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

Appendix 1

Instructions and screen shots (translated from Dutch)

Introduction

The experiment concerns the task of the judiciary. A judge tries cases that are brought before him. The verdict in a case is based on facts that parties put forward and inquiries that the judge conducts or orders. The cases that will be put to you are in the area of criminal law and concern the adjudication of criminal offences. The experiment deals only with the question whether the accused is guilty or not guilty (and not the determination of the punishment).

In the cases you will have to try, uncertainty exists about the culpability of the defendant. In practice, a judge must above all prevent that innocent defendants are condemned, not only in the interest of these innocent persons, but also because the real perpetrator will remain at large and may again commit crimes. At the same time the judge will want to prevent that real perpetrators are acquitted. The crime would remain unpunished and the perpetrator could commit new crimes. However, without sufficient evidence the accused must be acquitted: the charges have not been proven beyond reasonable doubt.

In the experiment you can earn points that will be exchanged to money at the end of the experiment: 100 points equals 1 euro.

The experiment consists of 30 cases. In each case you will be asked to take two decisions: Decision **A** concerns the acquittal or conviction of the accused; decision **B** asks you to estimate the probability that the accused is guilty.

Decision A: to acquit or to convict

In every case you will have to choose between conviction and acquittal. If your judgement is correct, thus either if you convict a real perpetrator or if you acquit an innocent person, you will earn 100 points. If you convict an innocent defendant, that will cost you 1500 points. If you acquit a real perpetrator, that will cost you 300 points. In addition you will receive a fixed salary of 100 points per period.

To summarize:

		Real situation: the accused is	
		the perpetrator	**innocent**
Your decision	**Conviction**	100	-1500
	Acquittal	-300	100

Decision B: the probability that the accused is guilty

In each case you will be asked to estimate the probability that the accused is guilty. With this answer you can earn points as well. See the separate sheet. For example: when you report a probability of 20% that the accused is guilty, you will earn 36 points if the accused is the perpetrator and 96 points if the accused is innocent. The table is constructed in such a way that it is to your advantage to give your opinion truthfully. If you click here[10] you will see an explanation and for the mathematically

[10] In a pop up window the following text appeared: "Example

inclined a formal proof. It is not necessary that you understand this proof; it is sufficient that you know it is in your own interest to provide your real belief.

You can only continue with the next case, when you have taken decisions A and B. You will be paid either for your decision A or your decision B. The computer determines at random which one of the two decisions will be paid out. It is advisable to make both decisions as best as you can, because you do not know which decision will be paid.

Evidence

A judge bases the decision to acquit or convict a defendant accused on the available evidence. Evidence can be incriminating or exonerating. Incriminating is for instance when a witness has seen the accused close to the scene of the crime around the time it took place. Exonerating is for instance when a witness has seen the accused far away from the scene of the crime at the time it occurred.

Incriminating and exonerating information that can be derived from a single piece of evidence can differ in strength of evidence. For instance, assume that in the fist of the victim of a violent crime blond hair that in all likelihood belongs to the perpetrator is found. If the accused has blond hair, this is incriminating evidence, but it is not very informative, because a large part of the population has blond hair. If the accused has black hair, the evidence is exonerating and much more informative. Furthermore, pieces of evidence can differ in strength. The DNA of the hair would offer much stronger evidence than the colour, as only a small number of people would have the same DNA-profile.

In the experiment you will receive for each piece of evidence information about the probability that this evidence will be happened upon in case of a perpetrator and the probability that this evidence will be happened upon in case of a person who is innocent. The evidence is stronger, the larger the difference between the two probabilities. The evidence will not be further described. Also, you will not be informed whether the evidence is put forward by the prosecution (public prosecutor) or the defense (lawyer of the accused).

In each case the computer determines randomly (probability of 50%) whether the accused has committed the crime or not. Next, the computer determines the

---

We will illustrate this with an example. You believe that the probability that the suspect is guilty is 70%. If you report 70%, you will earn in about 70% of the cases 91 points and in 30% of the cases 51 point, which is on average 0.70*91+0.30*51=63.7+15.3=89 points. If you do not report your real belief, but for example 90% you will earn more if the suspect is guilty (99 instead of 91 points) but this does compensate for the cases where the suspect is innocent (only 19 points): the average earnings will be 0.70*99+0.30*19=69.3+5.7=75 points, which is less than the 89 points you would earn with your honest report. The same holds when you report a lower probability, say 60%. In that case your average earnings will be 0.70*84+0.30*64=58.8+19.2=78 points. To conclude: it is in your own interest to report your true beliefs!

For the mathematically inclined we also provide a formal proof. You do not have to understand this proof to be successful in the experiment, but you should keep in mind that it is in your own interest to report your true beliefs.

The table is based upon the following formula. If p is the reported probability, the earnings are 50+p-[p*p+(1-p)*(1-p)]/2 if the suspect is guilty and 150-p-[p*p+(1-p)*(1-p)]/2 if the suspect is innocent. Assume that the real probability is q. Which p will optimize your expected earnings? The expected earnings are q*(50+p-[p*p+(1-p)*(1-p)]/2) +(1-q)(150-p-[p*p+(1-p)*(1-p)]/2) and to calculate the optimum we differentiate to p:

q*(1-p+(1-p))+(1-q)*(-1-p+1-p)=q*(2-2p)+(1-q)*(-2p)=2q-2pq-2p+2pq=2q-2p

This will be 0 only if p=q and this is a maximum (because the second derivation is -2 which is smaller than 0). QED”

corresponding items of evidence, using the relevant probability distributions (see below). This happens at random as well.

All possible inquires that lead to evidence fall in three categories. Several inquiries of each type can take place, possibly with contradictory outcomes. Each type of inquiry results in either an incriminating or an exonerating piece of evidence. All inquiries lead to evidence; the probabilities of incriminating or exonerating evidence add to 100%, both for the perpetrator and an innocent suspect. The strength of evidence of the three types differs.

Type of inquiry	Possible outcome	Code in experiment	Probability of evidence if the accused is the perpetrator	Probability of evidence if the accused is innocent	The strength of the evidence is found by dividing both probabilities
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

An incriminating piece of evidence has a strength that is larger than 1 and is more informative the larger the strength. An exonerating item of evidence has a strength that is smaller than 1 and is more informative the smaller the strength. For ease of exposition the items of evidence have a colour and font size. Incriminating evidence is given in red and exonerating evidence in blue. Font size varies with the strength of evidence. The strongest incriminating evidence is 2INC ; the probability associated with this evidence is 4 times as large when the accused is guilty than when the accused is innocent. The strongest exonerating evidence is 1EXO ; the probability associated with this evidence is 4 times as large when the accused is innocent than when the accused is the perpetrator. The one but strongest incriminating and exonerating evidence (1INC and 2EXO, respectively) are presented in a smaller font. Finally, the results of inquiry 3 are represented smallest: 3INC and 3EXO .

Procedure per case

The computer generates a case by determining at random whether the accused has committed the crime or not. These two possibilities have equal probability (50%). Of course, you will not be informed about the outcome.
Next, the computer generates a number of items of evidence. This number is not the same in each period, and varies at random between 3 and 6 (independent of the guilt

or absence of guilt of the accused). The computer generates the items of evidence in the following way:
1. the type of inquiry is chosen (40% probability of type 3 and 30% probability of type 1 and 2 each);
2. the outcome of the inquiry is determined, using the relevant probabilities.

For example: in this period the accused is the perpetrator. The computer decides randomly to do inquiry 2. This means that incriminating evidence will be generated with probability 64%. The computer draws a number between 1 and 100. Assume that the number is 74: because 74 is larger than 64 the evidence will be exonerating 2EXO with evidence strength 0.43.

All evidence will be presented at once, sorted on kind (coulor) and strength (size of font).

We will ask you some questions to check understanding and after that you will play some practice periods. Raise your hand if you need help.

Questions (participants could only continue after they answered all questions correctly. Feedback was provided to all questions by the computer, summarizing relevant parts of the instructions)
To make sure you understand the instructions we will ask you some questions.
Question 1.
Based upon the evidence, you decide to convict the accused (decision A). However, it turns out the accused was innocent. Assuming that in this period decision A is paid out, what will be your earnings (not including your salary)?
Question 2
Based upon the available evidence a participant believes the probability that the accused is guilty to be 75% (decision B). It turns out that the accused was the perpetrator. Assuming that decision B will be paid out, what will be her earnings (not including the salary)?
Question 3.
Assume that in the 30 periods exactly 15 accused are innocent and 15 are the perpetrator.

- A participant acquits all accused. What would this participant earn (assuming that decision A will be paid in all periods, excluding salary)?
- A participant convicts all accused. What would this participant earn (assuming that decision A will be paid in all periods, excluding salary)?
- A participant acquits all innocents and convicts all perpetrators. What would this participant earn (assuming that decision A will be paid in all periods, excluding salary)?

Question 4.
Is the following statement true or false? "Only one inquiry of type 2 can be done in a case."

Practice periods
(The participants played individually 6 practice periods, with feedback)

Trial by a panel of judges

Criminal cases in which the prosecutor demands a sentence of more than one year in prison are not judged by one, but by **three judges**. After having learned all the facts,

the judges discuss the case before coming to a decision. They have to decide collectively to convict or acquit the defendant.
The computer has randomly formed groups of three participants. These groups stay the same for the whole experiment. In each of the 30 cases you first consider the evidence individually, and also individually you make your decisions A and B. After that the discussion starts with the two other participants in your group. This discussion is by use of a computer chat box. You can take the time you need, and there is no reason to hurry. Just like in the real deliberation of judges, every member in the group has to give his or her opinion in this discussion. The collective decision can only be made after each participant has made at least one contribution to the discussion. Of course, longer discussions are possible.
If you feel that for you the discussion has ended, you can leave the chat box and make your decisions A and B. Your decisions can be the same as your primary decisions if you are not convinced by the arguments of the others, but it can also differ if your insights have changed. The point is to make the best possible decision. Better decisions lead to higher earnings.

Decision A

Also the other two members make a decision. The decision A of the panel is determined by the computer based upon these three decisions: if two or three members decide to acquit, the decision of the panel is to acquit. If two or three members decide to convict, the decision of the panel is to convict.

Decision B

The decision B of the panel (the estimation of the probability of guilt) is the **median** of the three decisions of the members.

> Example 1: the decisions B of the members are 20%, 52% and 70%: the decision B of the chamber is 52%.
> Example 2: the decisions B of the members are 10%, 42% and 43%: the decision of the chamber is 42%.

This procedure makes sure that if you believe the probability of guilt to be lower (or higher) than the other two members, you **cannot** influence the decision of the panel by filling in a lower (or higher) probability than your real opinion. If, in the first example above, the member who filled in 20% had filled in 0% the decision of the panel would have been the same. Likewise, the member who filled in 70% would not have changed the decision of the chamber by filling in 100%; the median and thus the decision of the panel would have stayed 52%. In the discussion you can try to convince the other members with arguments, but when the discussion has ended you can best fill in your real opinion.
You can earn money with decisions A and B; both the individual decisions and decisions of the panel.
We will start with the real cases in a minute. The differences with the practice cases are as follows.

- You make individual decisions as well as decisions with your group of three (the judicial panel);
- These periods will be paid out, either your individual decision A, your individual decision B, the panel decision A or the panel decision B. Which one of these four decisions will be paid out is determined randomly by the computer.
- Only at the end of the experiment you will learn for each period whether the

defendant is innocent or the perpetrator, and which of the four decisions will be paid out in each round.

We ask you not to disclose your identity or table number during the discussion in the panel.
If you raise your hand you will get a paper version of these instructions.

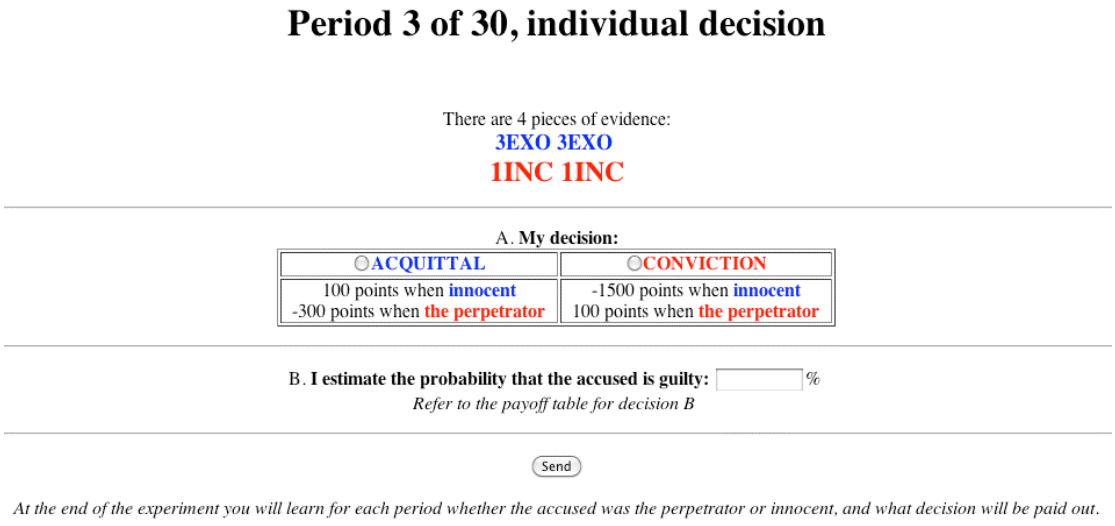
To make sure you understand the instructions we will ask you a question.

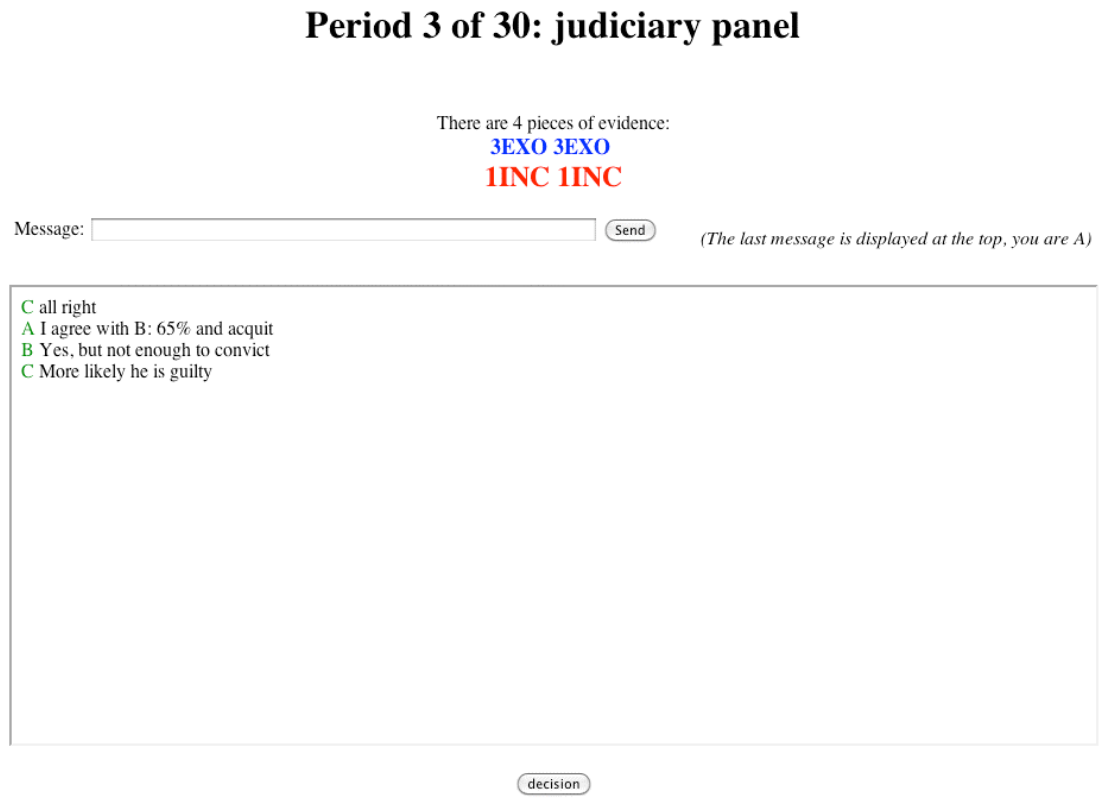
After the discussion in the chamber the three members have made the following decisions:

Member	Decision A	Decision B
A	ACQUITTAL	50%
B	ACQUITTAL	63%
C	CONVICTION	90%

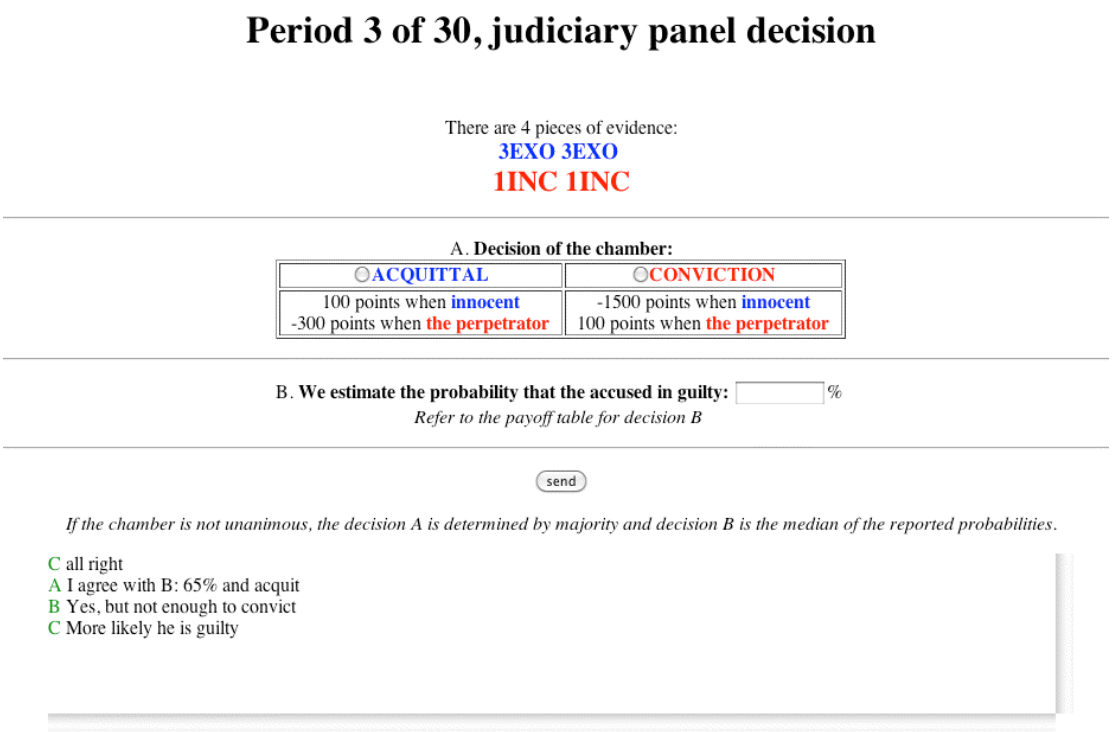
What is the decision of the chamber?
Decision A: ACQUITTAL / CONVICTION
Decision B: ….%

Examples of screens





(The decision button at the bottom of the screen only appears after all three members have contributed to the discussion).



(while making their decision the participant can read back the whole discussion)

Earnings decision B

Reported probability of guilt	Earnings if the accused is perpetrator	Earnings if the accused is innocent	Reported probability of guilt	Earnings if the accused is perpetrator	Earnings if the accused is innocent
0%	0.00	100.00			
1%	1.99	99.99	51%	75.99	73.99
2%	3.96	99.96	52%	76.96	72.96
3%	5.91	99.91	53%	77.91	71.91
4%	7.84	99.84	54%	78.84	70.84
5%	9.75	99.75	55%	79.75	69.75
6%	11.64	99.64	56%	80.64	68.64
7%	13.51	99.51	57%	81.51	67.51
8%	15.36	99.36	58%	82.36	66.36
9%	17.19	99.19	59%	83.19	65.19
10%	19.00	99.00	60%	84.00	64.00
11%	20.79	98.79	61%	84.79	62.79
12%	22.56	98.56	62%	85.56	61.56
13%	24.31	98.31	63%	86.31	60.31
14%	26.04	98.04	64%	87.04	59.04
15%	27.75	97.75	65%	87.75	57.75
16%	29.44	97.44	66%	88.44	56.44
17%	31.11	97.11	67%	89.11	55.11
18%	32.76	96.76	68%	89.76	53.76
19%	34.39	96.39	69%	90.39	52.39
20%	36.00	96.00	70%	91.00	51.00
21%	37.59	95.59	71%	91.59	49.59
22%	39.16	95.16	72%	92.16	48.16
23%	40.71	94.71	73%	92.71	46.71
24%	42.24	94.24	74%	93.24	45.24
25%	43.75	93.75	75%	93.75	43.75
26%	45.24	93.24	76%	94.24	42.24
27%	46.71	92.71	77%	94.71	40.71
28%	48.16	92.16	78%	95.16	39.16
29%	49.59	91.59	79%	95.59	37.59
30%	51.00	91.00	80%	96.00	36.00
31%	52.39	90.39	81%	96.39	34.39
32%	53.76	89.76	82%	96.76	32.76
33%	55.11	89.11	83%	97.11	31.11
34%	56.44	88.44	84%	97.44	29.44
35%	57.75	87.75	85%	97.75	27.75
36%	59.04	87.04	86%	98.04	26.04
37%	60.31	86.31	87%	98.31	24.31
38%	61.56	85.56	88%	98.56	22.56
39%	62.79	84.79	89%	98.79	20.79
40%	64.00	84.00	90%	99.00	19.00
41%	65.19	83.19	91%	99.19	17.19
42%	66.36	82.36	92%	99.36	15.36
43%	67.51	81.51	93%	99.51	13.51
44%	68.64	80.64	94%	99.64	11.64
45%	69.75	79.75	95%	99.75	9.75
46%	70.84	78.84	96%	99.84	7.84
47%	71.91	77.91	97%	99.91	5.91
48%	72.96	76.96	98%	99.96	3.96
49%	73.99	75.99	99%	99.99	1.99
50%	75.00	75.00	100%	100.00	0.00

Appendix 2

Description of cases

		Evidence						Decision	
period	**#evidence**	**EXO 1**	**EXO 2**	**EXO 3**	**INC 2**	**INC 1**	**INC 3**	**Objective probability**	**optimal decision**
1	6	3	1	1	0	1	0	1.0%	acquit
2	6	1	0	1	0	1	3	56.5%	acquit
3	4	0	0	2	0	2	0	70.3%	acquit
4	6	0	0	0	1	2	3	98.7%	convict
5	5	0	2	2	1	0	0	24.4%	acquit
6	6	1	3	2	0	0	0	0.9%	acquit
7	4	0	0	1	0	2	1	84.3%	convict
8	3	0	0	0	1	1	1	93.3%	convict
9	3	1	1	0	0	1	0	20.0%	acquit
10	5	1	1	1	0	0	2	13.8%	acquit
11	4	1	0	1	0	2	0	47.3%	acquit
12	3	2	1	0	0	0	0	2.6%	acquit
13	5	0	1	1	2	0	1	87.2%	convict
14	5	2	1	1	0	1	0	4.0%	acquit
15	4	0	3	0	1	0	0	24.1%	acquit
16	5	0	1	2	0	1	1	39.6%	acquit
17	3	1	0	1	0	0	1	19.8%	acquit
18	3	0	1	0	0	2	0	70.0%	acquit
19	3	0	1	1	1	0	0	53.2%	acquit
20	5	0	1	1	0	1	2	59.8%	acquit
21	5	0	0	2	1	1	1	85.9%	convict
22	3	0	1	0	0	0	2	49.2%	acquit
23	4	0	0	0	1	3	0	98.1%	convict
24	3	0	0	0	1	1	1	93.3%	convict
25	5	0	0	1	1	2	1	95.6%	convict
26	4	0	1	1	0	2	0	60.6%	acquit
27	4	1	0	2	0	1	0	20.2%	acquit
28	4	1	0	2	0	0	1	14.0%	acquit
29	4	1	1	0	1	0	1	39.2%	acquit
30	6	0	2	0	0	3	1	77.8%	acquit
Average								50.2%	26.7%

Table App2: Evidence, objective and subjective probabilities and optimal and actual decisions in part 1 per period.