Glöckner, Andreas; Engel, Christoph

Working Paper

Can we trust intuitive jurors? An experimental analysis

Preprints of the Max Planck Institute for Research on Collective Goods, No. 2008,36

Provided in Cooperation with:
Max Planck Institute for Research on Collective Goods

Suggested Citation: Glöckner, Andreas; Engel, Christoph (2008) : Can we trust intuitive jurors? An experimental analysis, Preprints of the Max Planck Institute for Research on Collective Goods, No. 2008,36, Max Planck Institute for Research on Collective Goods, Bonn

This Version is available at:
https://hdl.handle.net/10419/32193

Preprints of the
Max Planck Institute for
Research on Collective Goods
Bonn 2008/36

Can We Trust Intuitive Jurors? An Experimental Analysis

Andreas Glöckner
Christoph Engel


Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2008/36

Can We Trust Intuitive Jurors?
An Experimental Analysis

Andreas Glöckner / Christoph Engel

October 2008

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

Can We Trust Intuitive Jurors?
An Experimental Analysis

Andreas Glöckner / Christoph Engel

Abstract

Jury members do not normally have the privilege of a complete, unbiased picture of the case. To make the best of patently incomplete evidence, they cannot but at least partially rely on their intuition. We provide evidence for this claim based on self-report data as well as more subtle measures of unconscious modifications of the evidence in order to fit the favoured interpretation (coherence shifts). In three experiments we investigated whether members of a mock jury apply standards of proof in a normatively appropriate way, how well they take into account explicitly stated probability information, and which factors influence the size of coherence shifts. We found a mixed pattern of results: manipulation of the standard of proof influences conviction rates in the intended direction, but there are fewer convictions in both standard of proof conditions than normatively expected. When asked to indicate the minimum probability of guilt necessary for conviction, subjects do not sufficiently discriminate between "beyond a reasonable doubt" and "preponderance of the evidence". Even substantial manipulations of the posterior probability of guilt had very little effect on conviction rates. Reliance on intuitive processes seems to reduce the influence of explicitly stated probabilities. We furthermore found effects of verdict and of the probability manipulation on the size of coherence shifts. We argue that the performance of jury members could be improved by providing them with supplementary information on context, such that they are able to put explicit information on probabilities in perspective.

1. Introduction

Ever since there have been juries, there has been doubt about decision quality. Why should one believe that a bunch of laypeople is able to solve complex problems of inference? Not only, and of course by design, do they lack professional legal training. Legal procedure even exposes them to a structural asymmetry. The parties are represented by professional litigants who get a high premium for making the jury see the case in a light that is favourable to their cause. Jury decision-making might still be desirable on different grounds. Legal professionals might lack access to the full diversity of social reality. Decisions on behalf of the People might gain in legitimacy if they are actually taken by representatives of the public at large. Jury duty might be a way of turning passive subjects into active citizens. In this perspective, society trades a possible reduction in decision quality for these additional benefits.

Yet is the hypothetical trade-off real? Jurors are not supposed to handle legal doctrine. The difference in expertise is confined to assessing whether factual claims made by the parties are supported by the evidence. Jurors have to do this once, while professional members of the judiciary have a chance to gain experience. Does professionalisation matter? One might be tempted to draw an analogy to science. Most of hard science is about inference from probabilistic cues. Typically scientists are testing hypotheses derived from theory on already available or self-produced evidence. Obviously, jurors and professional judges are rarely in a position to rely on scientific techniques for drawing inferences, and even scholars who argue for greater openness of judicial procedure to probability theory are hesitant to impose mathematical training on the judiciary, let alone on jury members (for an overview see Jackson 1996). This is not a lack of sophistication in the legal discipline. In order to make powerful mathematical tools, such as Bayes' theorem, applicable, scientists must radically cut down on complexity. The judiciary does not have this luxury. It would not be permissible to "decontextualise" the case. This explains why, jurors and professional judges alike, must at least partially rely on their intuition (Day 1987; Guthrie, Rachlinski et al. 2007).

Judges and jury members do not act like miniature scientists that follow mathematical rules to calculate probabilities. It has been argued that their behaviour can be better explained by sense making (Pennington and Hastie 1991) and constructing coherent stories from the evidence. Decision-making seems to be often interpretative (Pennington and Hastie 1988). It relies on reasoning about the evidence, rather than an algebra like process (Pennington and Hastie 1988). Jurors attempt at creating a narrative story from the pieces of evidence they have heard (Pennington and Hastie 1986; Pennington and Hastie 1988; Pennington and Hastie 1993; Pennington and Hastie 1993) which can be considered as a mental model (Pennington and Hastie 1988); (also see Johnson-Laird 1983).

Story construction seems to be instantaneous. It starts early on, with hearing the first pieces of evidence (Hastie, Penrod et al. 1983; Pennington and Hastie 1988). It has been argued that jurors decide by matching stories to the representation of the verdict categories given to them in the judge's instructions on the law (Pennington and Hastie 1986). Whether a story is accepted, or

whether it is selected, depends on its goodness of fit (Pennington and Hastie 1993). In this assessment, jurors look out for coverage, coherence and uniqueness (Pennington and Hastie 1992). Coverage, coherence and uniqueness also determine the level of confidence (Pennington and Hastie 1988; Pennington and Hastie 1992).

There is mounting evidence on the character of the mental process which underlies story construction, and it has been argued that the same process underlies all forms of intuitive decision making (Glöckner and Betsch 2008a). In line with the basic claim from *Gestalt* psychology (Markus and Zajonc 1985), the assessment of the evidence often seems to be holistic (Simon 2004) and relies at least partially on an automatic process that has been developed from perception. It can be modelled by parallel constraint satisfaction (PCS) network models (Simon and Holyoak 2002). According to these connectionist models, decision-making progresses bidirectionally (Holyoak and Simon 1999; Simon, Krawczyk et al. 2004; Simon, Snow et al. 2004; Simon, Krawczyk et al. 2008). Not only do facts determine conclusions. Potential conclusions also affect the perception of the evidence. The mental model reconfigures itself until maximal coherence is achieved (Simon 2004).

The mechanism transforms the information input by automatically accentuating initial advantages for one or the other interpretation in the evidence. Over the consecutive iterations, information supporting the final decision is overestimated and conflicting evidence is underestimated. Information is thus polarised (Simon 2004). This process has been dubbed a coherence shift (Simon, Pham et al. 2001). It has been shown that coherence shifts can be pronounced (Simon 2004). Normally the construction of consistent interpretations is unconscious. Only the result is propelled back to awareness, for instance in the form of the feeling that one interpretation of the evidence is most appropriate ("I cannot prove it, but when I see it I know it"). In other cases feelings of sudden insight ("now I get it") or of unease ("something is fishy here") are produced. Note, however, that according to recent models (Glöckner and Betsch 2008) intuitive processes operate in close interaction with conscious deliberate processes of information search, information construction and supervision. Hence "deciding intuitively" does not necessarily mean that individuals do not deliberate before making a decision. Jury members will of course pay attention to the information presented, and they will consider it deliberately. Yet the core process of information integration: making sense of the information, forming a consistent interpretation which results in favouring an option, will be based on automatic processes. Conscious and unconscious pieces of information are considered and the information integration process remains opaque to the decider.

There is a rich empirical literature that compares judge and jury decision-making (Eisenberg, Hannaford-Agor et al. 2005; King and Noble 2005; Robbennolt 2005; Eisenberg, Hannaford-Agor et al. 2006; Read, Connolly et al. 2006; Eisenberg and Miller 2007; Spencer 2007). It has been argued that in many cases there are no striking differences in decision outcomes and quality (Robbennolt 2005). For instance, the ratio of compensatory to punitive damages is approximately the same in jury versus judge decision making (Eisenberg, Hannaford-Agor et al. 2006). In line with these findings, when the parties have a chance to replace jury trial by arbitration,

they rarely do so (Eisenberg and Miller 2007). They thus forego the opportunity to replace lay inference by professional inference. Interestingly, case complexity does not generate more disagreement between judges and jury members (Eisenberg, Hannaford-Agor et al. 2005).

All of this is converging evidence that, for making inferences, both judges and jurors rely on similar basic decision processes. Based on the findings on story construction, and given complexity in these decisions is high while human capacity for deliberate processing is bounded, it seems highly likely that both classes of decision-makers rely on intuition, which can be modelled by parallel constraint satisfaction. There is no need for upfront training to enable a decision-maker to use intuitive processes to integrate a multitude of pieces of information (Glöckner and Betsch 2008). This distinguishes intuition and scientific methods of inference, like Bayesian updating. If decisions can be based on general world knowledge ("is it plausible for a person to be a caring parent if she tortures her employees?"), no difference would be expected. On the other hand, decision making based on intuition in a specific domain ("is a case likely to break down if one piece of evidence is shown to be doctored?") does, of course, benefit from experience (cf. Dreyfus, Dreyfus et al. 1986).

While for many decisions there are no systematic differences between judges and juries, for specific classes of cases these differences are pronounced. In child abuse cases, juries are much more likely to convict (Read, Connolly et al. 2006). Juries are more frequently awarding punitive damages for financial injury, judges do so more frequently for bodily injury (Eisenberg, Hannaford-Agor et al. 2006). Juries are more likely to acquit if the defendant has no criminal record (Givelber and Farrell 2008). They are not sufficiently sensitive to selection bias in the presentation of evidence (Koehler and Thompson 2006) and not sufficiently likely to disregard inadmissible evidence (Steblay, Hosch et al. 2006). While rigorous proof would of course require experimental tests, it seems plausible that these differences can be put down to a property of intuitive decision-making. It is not under conscious control and therefore more liable to prejudice (Engel 2008) and emotional influences (Bright and Goodman-Delahunty 2006; Pettys 2007).

A questionnaire study on judge and jury agreement showed that judges also held different opinions about evidentiary strength. Consequently juries were more likely to convict when judges considered the evidence to be weak, and they were more likely to acquit when judges considered the evidence to be strong (Eisenberg, Hannaford-Agor et al. 2005). By design, jury members do not have the benefit of comparing across and learning from cases. This might explain the different consideration of evidence. Since they lack sufficient context and background knowledge, their judgement might be less valid. A lack of expertise could also explain why sentences are more severe if juries are not only responsible for assessing guilt, but also for sentencing (King and Noble 2005): Since juries cannot compare across cases, they are not in a position to put the case they have to decide in perspective.

While intuition is powerful in making the most of incomplete evidence and in instantly integrating huge amounts of information, its performance heavily relies on how well the mental model represents the underlying problem structure (Kahneman and Tversky 1972; Bar-Hillel 1980;

Barbey and Sloman 2007; Glöckner 2008). Besides many other influences, intuition is prone to fail when presented with specific numerical values, and in particular scientific mathematical argument, that are hard to evaluate because they are provided without context (cf. Hsee 1996; Finucane, Alhakami et al. 2000; Slovic, Finucane et al. 2002). Take, for instance, the information that an eye-witness is able to identify an accused person in 80% of the cases correctly. Jury members do not have the benefit of experience. Unlike professional judges, they cannot activate knowledge from comparable cases to evaluate this information and provide intuition with an appropriate mental representation. This might explain why juries perform poorly when it comes to using scientific results for assessing eyewitness reliability (Schmechel, O'Toole et al. 2006) and circumstantial evidence more generally (Heller 2006).

2. Research Question and Hypotheses

The fact that jury members make decisions based on intuitive processes and according to parallel constraint satisfaction models could have two normatively undesirable effects: First, standards of proof might be muted because of coherence shifts. A higher standard of proof might just lead to a stronger coherence shift. Conflicting evidence might just be devalued even more strongly. Thus, manipulations of the standard of proof might not influence conviction rates appropriately if individuals rely on intuition. Second, explicit information about the validity of a piece of evidence (i.e., the conditional probability of a stated fact, given this piece of evidence) might not get sufficient weight.

The first concern might be the downside of a key advantage of intuition. It enables the decision maker to come down on one side although the evidence is patently incomplete or the problem is visibly ill-defined. The parallel constraint satisfaction mechanism achieves this by spreading of activation. The evidence activates cues. These cues are positively or negatively related to decision options. If options exclude each other, they are negatively related. As a first step, the evidence provides support for at least some of the options. But the process does not stop here. Depending on how strongly an option is activated initially, it propels positive activation back to the supporting evidence, and negative activation to the conflicting evidence, and to the competing options. Based on the resulting re-assessment of the evidence, the process enters the next iteration, and so on. It is repeated as long as the marginal changes in the activation of the options are substantial. The option with the highest final activation is chosen, provided all competing options are sufficiently less activated and the network is overall sufficiently consistent (Glöckner 2008; Glöckner and Betsch 2008).

Courts are not allowed to refuse deciding a case. In principle, intuitive decision-making is therefore conducive to the goal of judicial procedure. However, decisions taken on incomplete grounds are by necessity error prone. Courts can err on both sides. They can convict the defendant although she is actually innocent. Or they can acquit the defendant although she is actually guilty. In criminal justice, the first risk of an alpha error (i.e., convicting an innocent person) is taken very seriously. Consequently the standard of proof is strict. The defendant may only be

convicted if she is guilty "beyond a reasonable doubt". In private lawsuits, however, many legal orders, and US law in particular, do not intend to privilege the defendant over the claimant. Beta errors (i.e., acquitting a guilty person) are taken as seriously as alpha errors. This translates into the more lenient standard of "preponderance of the evidence".

Intuition might be well equipped to apply the latter balanced standard. Yet one may wonder whether intuition is able to assess whether the strict standard of criminal procedure is met. In terms of mental mechanism, there are two related challenges. The law requires the decision maker to declare the evidence inconclusive although it would have been sufficient to find against the defendant under the preponderance of the evidence standard. For the different standards of proof to be effective, three conditions must hold. While the mental mechanism is designed to force a decision, it must be flexible enough to single out problems that are too hard to decide. The decision criterion (i.e. the required level of consistency) may not be the same all over. It must be possible to exogenously impose a stricter decision criterion on a class of problems.

An earlier experiment by Dan Simon seems to support these concerns. He had his subjects fulfil three consecutive tasks. He first asked them to rate the relevance of scrambled pieces of evidence. He then had them either convict or acquit the defendant, based on this evidence. Finally he had them rate the individual pieces of evidence a second time. As predicted by consistency maximisation models, subjects substantially re-rated the evidence such that it supported their individual decision (Figure 1). Those who had acquitted the defendant systematically deflated their agreement with statements containing inculpating evidence, and they inflated their agreement with statements containing exculpating evidence; this is how intuition forces decisions. However, those who convicted the defendant inflated exculpating evidence substantially more than acquitters deflated inculpating evidence (Simon 2004). While the difference was not statistically significant[1], it invites a hypothesis: subjects are likely to know that, in criminal procedure, convicting an innocent defendant is much graver than acquitting a guilty defendant. Instead of applying the stricter standard and acquitting the defendant, participants might have reacted by deflating exculpating evidence even more strongly.

1 Personal communication from Dan Simon.

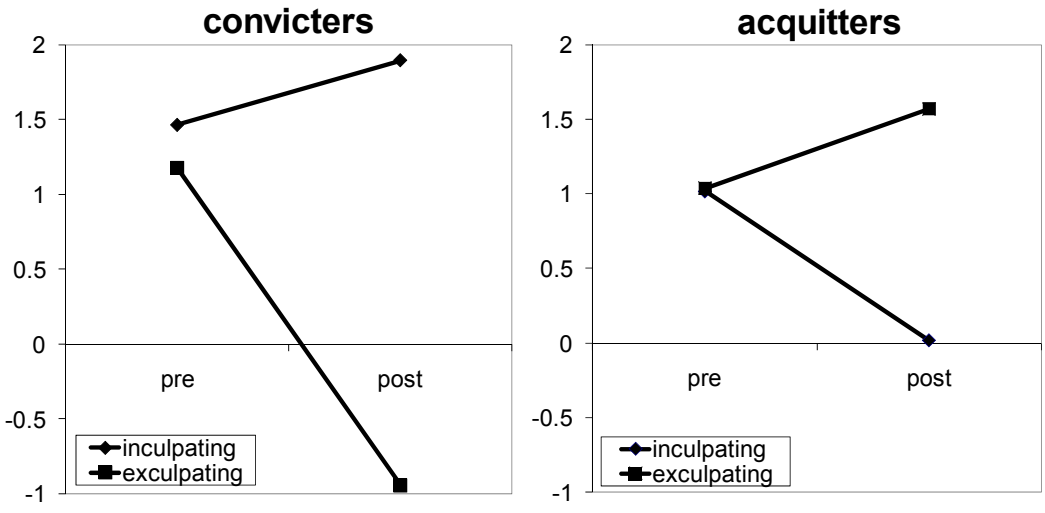

Figure 1
Dan Simon's Experiment: Coherence Shifts
mean rating pre / post decision
scale [-5,5]

The legal community is not interested in mental mechanism per se. In this paper, and in our hypotheses, we therefore primarily address behavioural effects, before we turn to the question whether (mock) jury members do indeed rely on intuition, and how this influences their decision.

H_1: strong version: Standards of proof do have no significant effect on conviction rates.

Even if we were able to reject the strong version of **H_1**, there might still be reason for normative concern. We therefore also test:

H_1: weaker version: Standards of proof have a smaller effect on conviction rates than would be normatively desirable.

For the weak version of **H_1**, the normative standard is hotly debated. The courts tend to focus exclusively on the alpha error. They thus exclusively aim at making false convictions sufficiently unlikely (leading case: *Addington v. Texas*, 441 U.S. 418, 422 (1979)). Scholarly critics call for a norm that takes also the corresponding beta error into account. They thus want the courts to establish a balance between false convictions and false acquittals (Kaye 1982; Orloff and Stedinger 1983; Lillquist 2002). While some scholars call for a quantification of the standard (Kagehiro and Stanton 1985; Kagehiro 1990; Saunders 2005; Tillers and Gottfried 2006; Weinstein and Dewsbury 2006), others are squarely opposed (Franklin 2006). If numbers are given, for "beyond a reasonable doubt" they are in the range of 90% to 99% certainty (Weinstein and Dewsbury 2006). If one quantifies "preponderance of the evidence", the probability of guilt must be (just) above 50%.

As explained above, lay persons' intuition might not adequately exploit scientific information that is provided in a numerical format without context. Specifically individuals might not be able to evaluate explicit information about the validity of information if no comparison standards are available. This leads to

H_2: strong version: Subjects are not sensitive to manipulations of explicit information on the validity of pieces of evidence.

Again, to be responsive to the normative discourse, we also consider

H2: weaker version: Explicit information on the validity of a piece of evidence has a lesser effect on conviction rates than would be normatively desirable.

For the weaker version of H_2, it is even more difficult to derive a standard from legal doctrine. On the European continent, the norm is holistic in the first place (Clermont and Sherwin 2002; Taruffo 2003). Professional and lay judges are asked to form a personal conviction, and to take on individual responsibility for the decision. In the US, many jury instructions for the "beyond a reasonable doubt" standard also have a discernible holistic flavour (see e.g. Pa. SSJI (Crim) 7.01), while those calling for quantification would want to turn judicial procedure in an exercise of updating priors in light of the evidence heard (Saunders 2005; Tillers and Gottfried 2006; Weinstein and Dewsbury 2006). Given these controversies, we do not have a numerical benchmark for the weak version of H_2.

We further investigate how subjects represent standards of proof, testing

H_3: When asked to quantify standards of proof, under the "beyond a reasonable doubt" standard subjects are willing to convict below a 90 % probability of guilt level.

H_4: When asked to quantify standards of proof, under the "preponderance of the evidence" standard subjects are willing to acquit although the probability of guilt is above 51%.

We test two psychological hypotheses regarding the underlying mental mechanism. This should allow evaluating our basic assumption that individuals partially rely on intuitive processes based on parallel constraint satisfaction (PCS). The first hypothesis addresses this question directly and uses self report measures. Competing models claim that individuals calculate deliberately (e.g., according to the rules of probability theory) or that they rely on (fast-and-frugal) heuristics. The term refers to radically simplified decision rules that do only take a small part of the information into account (e.g. Gigerenzer, Todd et al. 1999). We thus test

H_5: When they are asked to infer a fact from a set of conflicting pieces of evidence, participants describe their decision strategy as being in line with intuitive processing based on PCS.

The second hypothesis concerns the key behavioural difference for applying intuition predicted by PCS models but not by the competing models (i.e. the rational Bayesian model and heuristics):

H_6: If asked to re-rate the evidence after they have taken their decision, subjects increase the weight of evidence supporting their decision, and they decrease the weight of conflicting evidence.

While the "preponderance of the evidence" standard is balanced, the "beyond a reasonable doubt" standard asks jurors to treat the evidence differently if they decide to convict, compared to acquittal. Hence, one would expect that intuitive jurors only convict if they have sufficiently devaluated all contrary evidence. We therefore tested **H_7**:

H_7: Under the "beyond a reasonable doubt" instruction, coherence shifts are more pronounced if participants convict as compared to acquit.

3. Experimental Design

To test our hypotheses, we conducted a series of three experiments. In the first experiment we manipulated two standards of proof and two levels of conditional probability of guilt, given the evidence. The second and third experiments were meant to check the robustness of our findings from the first experiment. In the first experiment, we had increased the plausibility of the different instructions by framing the treatment with the "beyond a reasonable doubt" instruction as a criminal trial, and the treatment with the "preponderance of the evidence" instruction as arbitration. In the second experiment, as a robustness check, we paired both instructions with both frames. As will be reported in greater detail below, in the first two experiments we found that changes in stated conditional probabilities had very little effect on conviction rates. As an even stronger test of **H_2**, in the third experiment we increased the stated probability of guilt, given the evidence, to very high numbers. In the three studies we used the following treatments:

Experiment	Treatment	Standard of Proof	Frame	Probability	N
1	1	brd	criminal	medium	20
	2	brd	criminal	high	22
	3	poe	arbitration	medium	20
	4	poe	arbitration	high	22
2	1	brd	criminal	high	19
	2	brd	arbitration	high	20
	3	poe	criminal	high	18
	4	poe	arbitration	high	17
3	1	brd	criminal	high	43
	2	brd	criminal	very high	42

Table 1
Experimental Design
legend: brd beyond reasonable doubt, poe preponderance of the evidence

All experiments were conducted at the University of Erfurt. They consisted of two parts and a 20 minute filler task between the parts. In the first experiment, 84 students took part (22 male, 64 female; 18 to 29 years). In the second experiment, 74 students took part (25 male, 49 female; 17 to 30 years). In the third experiment, 87 students took part (31 male, 56 female; 17 to 36 years). All subjects received EUR 6 (approximately U.S.$7.50) for their participation. All experiments lasted about an hour.

In the first part of each experiment, subjects received pieces of evidence in a scrambled order, and without knowing that they would later have to pass judgement on the guilt of the defendant under one of the instructions. In the second part, subjects were first asked to take a decision. After they had decided, they were again asked to rate the evidence.

We used a translated and slightly modified version of a complex legal case constructed und repeatedly used by Dan Simon and colleagues (Holyoak and Simon 1999; Simon 2004; Simon, Snow et al. 2004); the complete case can be found in the appendix. In this case, a company accuses one of its employees of having stolen money from the company safe. The case consists of six pieces of information pro-guilty and contra-guilty, each. This information consists of facts and background beliefs. It is known that the money was stolen using the regular access code which only a few persons had. The money was stolen in the evening and the time was recorded. The crucial pro-guilty facts are a) the number of persons who knew the access code to the safe which was used to steal the money, b) the confidence level of an eyewitness who afterwards reported having seen the accused person at the site of crime, and c) the relative frequency of a certain type of car in the region which was seen at the site of crime and which is also driven by the accused person. The strongest contra-guilty fact is that d) the accused person was seen shortly after the date of crime in a place which was hard to reach in such a short time.

The manipulation of the factor Probability varied the number of persons who had the access code for the safe (medium: 18 persons, high: 8 persons, very high: 4 persons), the self-reported confidence level of the eyewitness (medium: 80% confident, high: 95% confident, very high: 99% confident), and the relative frequency of the type of car which was seen at the site of crime and is driven by the accused person (medium: 6%, high: 0.1%, very high: 0.01%).

To manipulate the instructions, we used translated versions of the official model jury instructions of the Ninth Circuit.[2] Our instructions read:

> "You should decide by a preponderance of the evidence, it means you must be persuaded by the evidence that the claim is more probably true than not true. You should base your decision on all of the evidence, regardless of which party presented it."
>
> "Please note that in criminal cases accused persons are particularly protected. They should only be convicted if the evidence is so convincing that there is no reasonable doubt that the person is guilty. Proof beyond a reasonable doubt is proof that leaves you firmly convinced that the defendant is guilty. It is not required to prove guilt be-

2 The instruction is available online at www.ce9.uscourts.gov (2003 ed.).

yond all possible doubt. A reasonable doubt is a doubt based upon reason and common sense and is not based purely on speculation. It may arise from a careful and impartial consideration of all the evidence, or from lack of evidence.

If after a careful and impartial consideration of all the evidence, you are not convinced beyond a reasonable doubt that the defendant is guilty, it is your duty to find the defendant not guilty. On the other hand, if after a careful and impartial consideration of all the evidence, you are convinced beyond a reasonable doubt that the defendant is guilty, it is your duty to find the defendant guilty."[3]

In the first part, subjects read short scenarios about social interactions. These scenarios contained the relevant cues of the legal case in different situations and were rated on a scale from -5 (*strongly disagree*) to 5 (*strongly agree*). For instance, participants read that an eyewitness was 80% confident of having identified a specific person bringing some flowers for a colleague after work. They then were asked how strongly they agree with the statement that the identification makes it likely that this person indeed brought the flowers.

In the second part, subjects were presented with case materials which consisted of a general instruction, including a standard of proof, some background information on the accused person, a summary of the evidence, the arguments of the company, the arguments of the defense, a sheet to indicate the decision and decision related information (see appendix). Individuals indicated their decision. They rated the confidence in their decision on a scale ranging from *completely uncertain* (0) to *completely certain* (10). They had to estimate the probability that the accused person had stolen the money from the safe, as well as the probability that another person took the money (all estimates in percent). Finally, they were asked to specify the level of probability necessary for convicting under the respective standard of proof. In a post-test, subjects re-rated the evidence, using the same scale as in the first part from -5 (*strongly disagree*) to 5 (*strongly agree*). They, for instance, indicated how strongly they agree with the claim that eyewitness identification with 80% certainty makes it likely that the accused person had stolen the money. At the end of the experiment, subjects filled out a questionnaire meant to capture mental mechanism by self-report, which we used to test $\mathbf{H_5}$.

For all hypotheses, we separately report results from experiment 1 and pooled results from all experiments. Results from experiment 2 in isolation are reported in the section on the standard of proof manipulation. Results from experiment 3 in isolation are reported in the section on the probability manipulation.

4. Results

We report results in two different ways. We start with descriptive statistics and inference-statistics on our hypotheses. Then we use a regression approach to analyze how our independent variables jointly influence our dependent variables.

3 Note, that in Experiment 2 in one condition participants applied the beyond reasonable doubt standard to an arbitration case. In this condition in the instruction, "criminal case" was replaced by "arbitration case".

a) Standards of Proof

According to the strong version of $\mathbf{H_1}$, conviction rates should not be significantly influenced by standards of proof. We observed higher conviction rates in the preponderance of evidence condition as compared to the beyond reasonable doubt condition (Table 2): Exp.1: $\chi^2(1, N=84) = 16.46$, $p < .001$, Exp. 2: $\chi^2(1, N=74) = 3.26$, $p = .071$. Hence, the strong version of $\mathbf{H_1}$ can be rejected. Although it is likely (and will be shown below) that individuals relied on intuitive consistency maximizing processes, the standard of proof influences conviction rates in the normatively intended direction.

						Conviction Rate		Coherence Shift		Subjective Probability		Subjective Norm		Confidence	
Exp	Treat	Stand	Frame	Prob	N	%	n	M	SD	M	SD	M	SD	M	SD
1	1	brd	crim	medium	20	.05	1	0.78	1.29	42.70	20.50	83.70	19.70	6.25	1.97
	2	brd	crim	high	22	.14	3	1.36	1.64	45.59	29.54	86.09	15.45	5.59	1.82
	3	poe	arb	medium	20	.40	8	1.07	1.18	52.75	21.33	72.55	14.83	5.55	1.47
	4	poe	arb	high	22	.59	13	2.17	2.04	55.91	27.19	75.27	17.58	5.81	2.20
2	1	brd	crim	high	19	.26	5	1.56	2.16	53.89	30.16	89.84	11.95	6.05	2.27
	2	brd	arb	high	20	.25	5	1.13	2.33	54.40	27.20	85.25	14.12	5.80	2.86
	3	poe	crim	high	18	.61	11	1.49	1.46	54.56	25.67	81.50	18.62	6.67	1.64
	4	poe	arb	high	17	.29	5	1.56	1.90	47.82	27.34	75.00	18.03	6.18	1.94
3	1	brd	crim	high	43	.19	8	0.82	1.73	59.07	25.57	87.84	14.61	5.81	2.44
	2	brd	crimi	very high	42	.24	10	1.44	1.36	49.14	26.88	81.07	17.02	5.93	2.49

Table 2
Descriptive Statistics

Conviction rate is the proportion of convictors. Coherence shifts represent the averaged size of reevaluation of evidence in line with the judgment from pre to post test, 0 indicating no changes and increasing coherence shifts with increasing values. Subjective probability indicates participants' rating of the probability that the accused person committed the offense. Subjective norm is the rating of the probability above which a person would in general be willing to convict. Confidence indicates the confidence rating, with higher numbers indicating higher confidence.

Experiment 2 was conducted to disentangle the naturally confounded effects of the standard of proof and the framing of a decision. Specifically, the criminal procedure versus arbitration frame was fully crossed with the standard of proof manipulation. If one splits the dataset according to frames, it turns out that the difference between standards of proof is insignificant in the arbitration frame, $\chi^2(1, N=37) = 0.09$, $p = .76$ but significant in the criminal case frame, $\chi^2(1, N=37) = 4.56$, $p = .033$. Note, however, that the sample size was rather small which might account for the non-significant effect in the arbitration frame.

For the weak version of $\mathbf{H_1}$, the probability manipulation matters. The manipulation varied the number of persons who had the access code for the safe (medium: 18 vs. high: 8 vs. very high: 4), the self-reported confidence level of the eyewitness (medium: 80% vs. high: 95% vs. very high: 99%), and the probability of the type of car which was seen at the site of crime and is driven by the accused person (medium: 6% vs. high: 0.1% vs. very high: 0.01%). This manipulation dramatically changes the posterior likelihood that the person committed the crime. If we

focus on these three quantified pieces of evidence, and if we estimate the probability that the person could have reached the far distant place in the short time available to be 25%, in our three probability treatments, the posterior probability that the accused person committed the crime rises from 55.13% over 99.89% to 99.999% (on the assumption that all probabilities are independent and that the prior probability of guilt is chance, i.e. .5) [4].

If these numbers are taken literally and assuming that, in the beyond a reasonable doubt condition, persons should be convicted if the probability of guilt given the evidence is higher than 90%, we would have to test the weak version of $\mathbf{H_1}$ by the hypothesis that all subjects acquit under the beyond a reasonable doubt standard and medium probability, and that all subjects convict under all other conditions. However the posterior probability of 55.13% is, of course, sensitive to our estimate of the probability of reaching the other location in time. If one reduces this estimate to 20%, the posterior probability of guilt is down to 48.0 %[5]. This would imply acquittal even under preponderance of the evidence. We therefore consider it more appropriate to expect about half of the subjects to convict under preponderance of the evidence and medium probability, and the other half to acquit. In the beyond reasonable doubt condition we would expect for the medium probability condition no decisions for guilty, in the high probability condition many guilty decisions, and in the very high condition almost all persons should decide for guilty.

We essentially found a good fit with this normative distribution in the medium probability condition. However, in all other conditions, particularly with high and very high probabilities, conviction rate was far below the normative standard, as shown by Table 3. From a normative point of view and taking into account probabilistic conventions about standards of proof, participants' conviction rate was far below the normative expected proportion, both in the beyond reasonable doubt condition and the preponderance of evidence condition.

4 We calculate the following way:

$$p(g \mid ev_1 \cap ev_2 \cap ev_3 \cap ev_4) = \frac{p(ev_1 \mid g) * p(ev_2 \mid g) * p(ev_3 \mid g) * p(ev_4 \mid g) * p(g)}{p(ev_1 \mid g) * p(ev_2 \mid g) * p(ev_3 \mid g) * p(ev_4 \mid g) * p(g) + p(ev_1 \mid i) * p(ev_2 \mid i) * p(ev_3 \mid i) * p(ev_4 \mid i) * (1 - (p(g)))}$$

where g stands for "guilty", i stands for "innocent", and ev_i stands for the four quantifiable pieces of evidence. Of course, $p(ev_i \mid g) = 1 - p(ev_i \mid i)$.

5 The posterior probability under high probability changes to 99.85 %. The posterior probability under very high probability is still approximated by 99.999 %.

experiment	probability manipulation	standard of proof	normatively expected probability of conviction	observed probability of conviction	p-value in binomial test against p=.50
1	medium	pre	0.5	0.4	0.5
		brd	0	0.05	<.001
	high	pre	1	0.59	0.26
		brd	1	0.14	<.001
2	high	pre	1	0.46	0.37
		brd	1	0.26	0.002
3	high	brd	1	0.19	<.001
	very high	brd	1	0.24	<.001

Table 3
Comparing Conviction Rates with the Legal Norm

Binomial Tests, H_0 p=.05. In exp 1 medium pre, the test is two-sided. All other tests are one-sided, determining the probability of observing a conviction rate that low if the true conviction rate is .5

Only in experiment 1, in the high probability + preponderance of the evidence condition, the conviction probability is above .5. In this condition, and in experiment 2 with preponderance of the evidence, there is no statistically significant deviation from judging randomly (p=.50). In all other treatments, the deviation is significant. Except for experiment 1, medium probability and beyond a reasonable doubt, this is a significant violation of the legal norm.

b) Explicit Information on Probabilities

The strong version of $\mathbf{H_2}$ implies that conviction rates are insensitive to increasing the stated conditional probability of guilt, given the evidence. Separately for experiment 1 and 3, we conducted χ^2-tests to compare conviction rates over probability manipulations. Both tests turned out not significant, $\chi^2(1, N=84) = 1.93$, $p = .17$, and $\chi^2(1, N=85) = 0.35$, $p = .56$. In both experiments there was at least a slight tendency to increased convictions with increasing probability (Exp 1: medium: 22.50 % vs. high: 34.88 %; Exp 3: high: 19.05 % vs. very high: 23.08 %, cf. Table 2 Table 2). Since the data support the strong version of $\mathbf{H_2}$, there is no need to consider the weak version.

The fact that individuals partially disregard explicit information on probabilities is in line with earlier findings showing that subjects rely on context information to interpret evidence (Hsee 1996). In our setting, probabilistic information was provided without context to evaluate it. For instance, the likelihood that the eye-witness is correct was manipulated from 80% to 99%, which might all be considered rather reliable, if one has no comparison standard. Our participants, like jury members, had not sufficient experience to evaluate probabilities by, for instance, comparing stated probabilities to probabilities in previous cases.

c) Subjective Norm

In $\mathbf{H_3}$ we expect subjects to represent the standard "beyond a reasonable doubt" in a normatively problematic way.

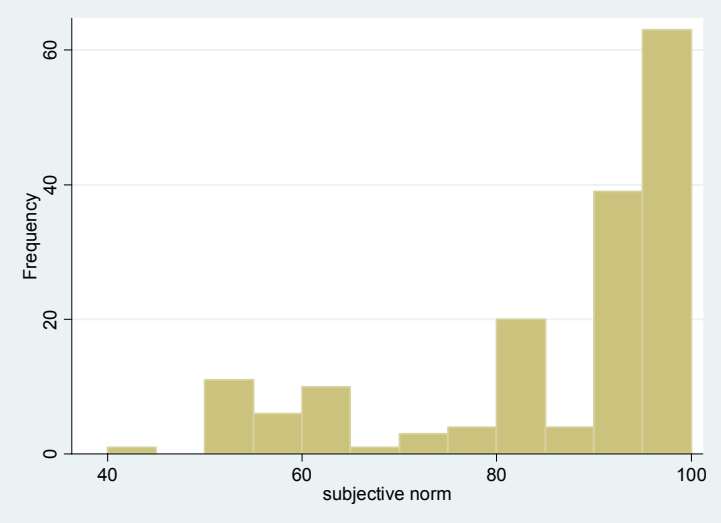

Figure 2
Subjective Norm under Beyond a Reasonable Doubt

The minimum probability of guilt accepted by the legal literature under the "beyond a reasonable doubt" standard is 90%. In our subject pool, the mean is 85.30. A *t*-test rejects the null hypothesis that the subjective norm is at or above 90% at $p<0.001$. In line with our expectations we found that the explicated standard for conviction was set too low. Note that conviction rates, at least in the high and very high probability conditions, were also too low. This indicates that although the explicated standard was set very low, this did not lead to too many convictions. Similar results concerning low levels of explicit standards have been found in the literature (Saunders 2005) (see also Bowers, Foglia et al. 2006).

Likewise in $\mathbf{H_4}$ we expect subjects to be too scrupulous under the "preponderance of the evidence" standard. The legal norm is a probability of guilt just above 50%.

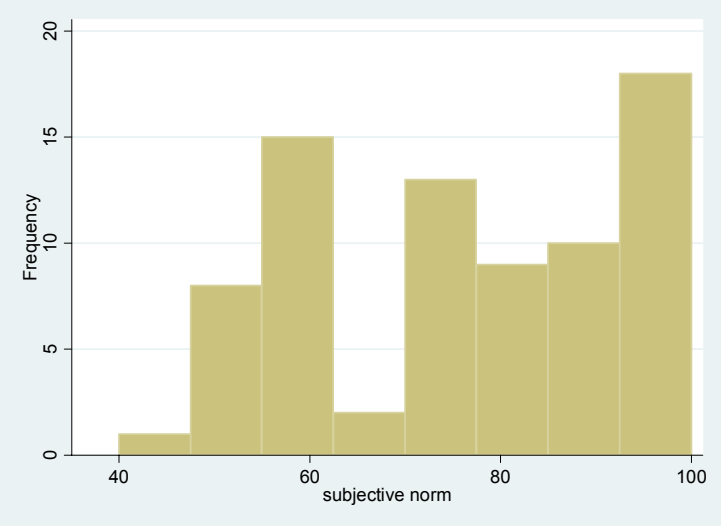

Figure 3
Subjective Norm under Preponderance of the Evidence

Here the deviation from the legal standard is even more striking. In our experiment, the mean is 75.96 %. A *t*-test rejects the null hypothesis that the subjective norm is at 51% at $p<0.001$. Note however that the German legal order has a standard of proof similar to "beyond a reasonable doubt" all over so that the "preponderance of the evidence" instruction might have been somewhat alien to our subjects.

d) Measures of Mental Mechanism

To test $\mathbf{H_5}$, at the end of Experiment 3 we asked our subjects to fill out a questionnaire. The questionnaire intended to measure whether individuals applied a) mathematical rational strategies (e.g., Bayes' Theorem) based on deliberate calculation (*rational strategy*), b) used simple rules of thumb (*heuristics*) which ignore most information, or c) in line with the PCS approach constructed consistent interpretations of the evidence and checked them for consistency using intuitive processes (*consistency maximizing*). It furthermore d) aimed to investigate if individuals are aware of coherence shifts (i.e., that they unconsciously change valuation of evidence). The questionnaire consisted of 13 statements which were all rated on a scale ranging from *strongly disagree* (-5) to *strongly agree* (+5). The statements and their connections to the research questions a) to d) are summarized in Table 4.

	Items	Strategies and Predictions				Results			
		Rational Strategy	Heuristic	Consistency Maximizing	Awareness of Coherence Shifts	*M*	*SE*	Agree (H1:μ>0)	Not Agree (H1:μ<0)
1)	I used all available information in my decision.	yes	no	yes		3.69	0.14	accept	
2)	I applied mathematical formulas to calculate the probability that H. is guilty.	yes	no	no		-3.76	0.22		accept
3)	I first tried to detect the most reliable piece of information and then I decided solely on the basis of this information.	no	yes	no		0.35	0.36		
4)	I counted the number of arguments for and against H. and decided solely on the basis of this amount.	no	yes	no		-2.75	0.28		accept
5)	In my decision, I took into account the reliability of all arguments for and against H.	yes	no	yes		3.46	0.15	accept	
6)	I estimated the reliability of the arguments for and against H., added them up and compared the results.	yes	no	no		-0.83	0.32		accept
7)	I checked the overall-argumentation for and against H. for consistency.			yes		2.99	0.21	accept	
8)	I tried to find the best possible interpretation of the provided information.			yes		2.60	0.22	accept	
9)	My judgment of the evidence was influenced by the interpretation of the circumstances.				yes	0.62	0.28	accept	
10)	My decision was based on an objective consideration of the information.	yes			no	2.18	0.25	accept	
11)	Under other circumstances, I would judge the reliability of the provided information similarly.				no	1.49	0.26	accept	
12)	For myself, I set a specific numerical probability limit, above which I convicted a person.	yes				-0.74	0.37		
13)	In my decision I have checked whether the circumstances taken together let me come to the firm conviction that the accused person is guilty.	yes		yes		3.20	0.21	accept	
Number of Hypotheses accepted		5	2	8	1				
Number of Hypotheses significantly rejected		2	3	0	2				
Insignificant Results / H0 retained		2	1	1	0				

Table 4
Questionnaire

The hypotheses concerning the four research questions (a-d) were analyzed by conducting *t*-tests against the null hypothesis that persons are undecided concerning the statement (H_0: $\mu=0$) testing both directed alternative hypotheses (Agree with the statement H_1: $\mu>0$; Disagree with the statement H_1: $\mu<0$) using an alpha level of 5% (two tailed). The numbers of significantly supported and rejected hypotheses, as well as the number of non-significant results are shown in the last three rows of Table 4. The results indicate a partial agreement but also significant disagreement with the hypotheses derived from a rational strategy. There was a majority disagreement with the statements derived from heuristics models. All but one statements derived from the PCS approach (consistency maximizing) could be supported by the data. And finally, individuals seem to be partially aware of the fact that their judgment of the evidence is influenced by the circumstances but they also strongly believe that they use information in a rational manner.

Our evidence therefore supports the view that in line with the predictions of the PCS approach individuals relied in their decisions on intuition that was based on automatic unconscious consistency maximizing. There is counterevidence against the two competing hypotheses: deliberate decision making, and the use of simple heuristics.

e) Coherence Shifts

In order to test $\mathbf{H_6}$ which states that individuals systematically modify the evaluation of evidence in the decision process (i.e. that they exhibit coherence shifts), we calculated for the six pro-guilty and contra-guilty pieces of evidence, each, difference scores between the rating in the pre-test and the post-test. These difference scores were averaged and constituted the variables RerateInnocent and RerateGuilty. The former measured coherence shifts for the contra-guilty arguments and the latter for the pro-guilty arguments. In both cases, positive values indicate that the subject attaches greater weight to these arguments after the decision than before it. Likewise, a negative value indicates that the subject has devalued the respective class of arguments.

Participants consistently, and strongly, devalue the weight of conflicting evidence, as indicated by the large negative values in Figure 4. They also increase the weight of supporting evidence, although to a lesser extent.

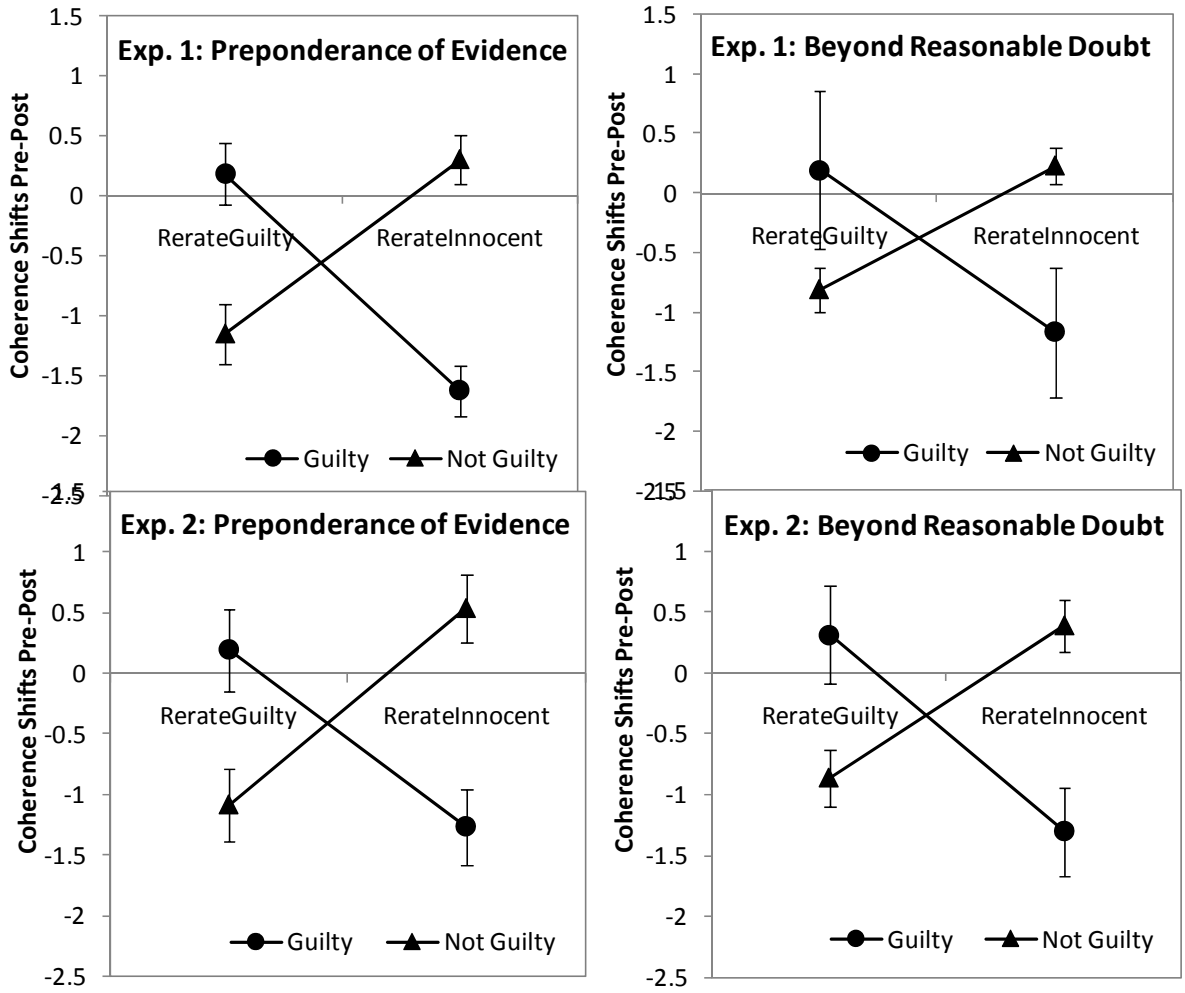

Figure 4
Coherence Shifts in Experiments 1 to 3

Positive numbers indicate an increase in the valuation of the information from pre- to post-test. Error bars indicate standard errors of the means

To test whether these coherence shifts are statistically significant, we conducted a repeated measurement analysis of variance (ANOVA) and expected a significant interaction effect between the verdict and the rerating of the evidence (RerateGuilty vs. RerateInnocent). To do so we recoded the data on coherence shifts such that each subject is tested twice on its rerating behaviour, once for inculpating and once for exculpating evidence, resulting in a within subject factor. In experiment 1, verdict, standard of proof, frame and probability are between subjects factors. The interaction between verdict and the within subjects factor is significant, $F(1, 76) = 26.49$, $p < .001$.[6] In experiment 2, verdict, standard of proof and frame are between subjects factor. The interaction between verdict and the within subjects factor is again significant, $F(1, 66) = 32.26$, $p < .001$. In experiment 3, verdict and probability are between subjects factors. Again the targeted interaction is significant, $F(1, 81) = 48.58$, $p < .001$. Hence, we found significant coherence shifts in all three experiments and rejected H_6.

We further aimed to investigate whether, in the beyond reasonable doubt condition, there were stronger coherence shifts if participants convicted as compared to acquitted (**H_7**). As can be seen by the steeper slope of the lines for convictors compared to acquitters in the beyond reasonable doubt condition (Figure 4), **H_7** was descriptively supported by the data. Statistical tests were conducted on the overall data set and hence are reported in the summary analysis section.

5. Summary Analyses

We conducted two regressions to analyze the results of all three experiments, controlling for all manipulated factors at the same time.

First, we analyzed the individuals' decisions by a logistic regression, with choices (0…innocent, 1…guilty) as criterion and our manipulated factors as predictors: probability coded in two dichotomous variables (p1 comparing medium against high and very high [-1, 0.5, 0.5] and p2 comparing high against very high [0, -1, 1]), framing (arbitration 0, criminal case 1), standard of proof (preponderance of the evidence 0, beyond reasonable doubt 1). Results support the findings reported above (Table 5). The standard of proof instruction had a strong effect on conviction rates (an odds ratio close to 0 implies a strong negative effect), whereas framing had no effect. Overall there is a marginally significant effect for the manipulation from medium compared to high and very high probability (p1), pointing into the expected direction: with higher probability, there are more convictions. There was essentially no effect of the manipulation from high to very high probability (p2).

6 All the repeated measures ANOVA are complete models. They thus include all possible interaction terms.

```
Logistic regression                                     Number of obs   =        243
                                                        Wald chi2(4)    =      24.30
                                                        Prob > chi2     =     0.0001
Log pseudolikelihood = -132.83734                       Pseudo R2       =     0.0838
------------------------------------------------------------------------------
                 |               Robust
     Convict     | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
Standard of Proof |   4.856548   2.071889     3.70   0.000     2.104699    11.20638
     Framing     |    .941505   .4079239    -0.14   0.889     .4027396    2.201004
          p1     |   .5636294   .1694421    -1.91   0.056     .3126797    1.015986
          p2     |   .8758083   .1921611    -0.60   0.546     .5697009    1.346391
------------------------------------------------------------------------------
```

Table 5
Logistic Regression Explaining Conviction Rates
reporting odds ratios and using robust standard errors

Second, we analyzed the size of coherence shifts by conducting a linear regression with coherence shifts as dependent variable, using robust standard errors. The variable coherence shift was computed conditional on the verdict so that a positive number indicates that the evidence is rerated to support the verdict. Specifically, for individuals who convicted coherence shift was computed by subtracting RerateInnocent from RerateGuilty, and vice versa for persons who acquitted. Predictors were again the factors convict, probability (i.e., the contrasts p1 and p2), framing and standard of proof.

```
Linear regression                               Number of obs =     243
                                                F(  5,   237) =    2.81
                                                Prob > F      =  0.0173
                                                R-squared     =  0.0435
                                                Root MSE      =  1.7088

------------------------------------------------------------------------------
coherence    |               Robust
shift        |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
     convict |   .4472515   .2512655     1.78   0.076                .1168775
    standard |  -.3479402   .3487487    -1.00   0.319               -.0938175
       frame |  -.1312495   .3510604    -0.37   0.709               -.0356298
          p1 |   .4049031   .1747711     2.32   0.021                .130527
          p2 |   .1461571   .1362051     1.07   0.284                .0653693
       _cons |   2.395058   .5001954     4.79   0.000                       .
------------------------------------------------------------------------------
```

Table 6
Linear Regression Explaining Coherence Shifts

The effect of convict is marginally significant, indicating that there are stronger coherence shifts if participants convict. Separate regressions for the two standards of proof reveal that the effect of convict is entirely driven by an effect in the beyond reasonable doubt condition (beta=.15, p=.04) and that there is no effect in the preponderance of evidence condition (beta=.06, p=.60). Hence, our data support $\mathbf{H_7}$ that, in the beyond reasonable doubt condition, there are stronger coherence shifts if persons convict, as compared to acquit.

Seemingly, from a normative perspective this is troublesome news. Participants react to the stricter instruction by stronger coherence shifts. Yet closer inspection of the data reveals that, on the contrary, the effect is driven by normatively desirable behavior.

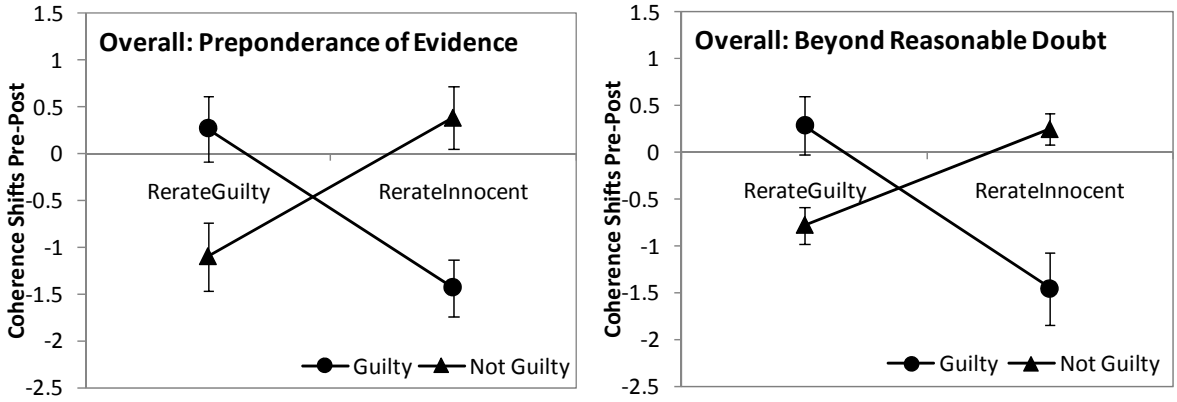

Figure 5
Rerating of Evidence by Standard of Proof

It would be normatively problematic if subjects neutralised the stricter instruction by devaluing exculpating evidence more strongly. Yet, as Figure 5 indicates, there is no significant difference between the handling of conflicting evidence if subjects convict (in a regression with the rerating of conflicting evidence as the dependent variable, frame, p1 and p2 as controls, and robust standard errors, the regressor for standard of proof is negative and insignificant, beta=-.17, p=.39). However there is a (weakly significant) difference if participants acquit (beta=.18, p=.09). If they acquit, under the beyond a reasonable doubt instruction, participants devalue inculpating evidence less strongly than under the preponderance of the evidence instruction. This is exactly what the legal order wants. Under the stricter standard of proof, many participants acquit although they believe the inculpating evidence to be strong, yet not strong enough. Apparently, many participants acquit "for want of evidence".

In the regressions of Table 6, the effect of p1 is significant, whereas p2 has no discernible effect. Hence, coherence shifts are stronger in the high and very high probability condition, as compared to the medium probability condition. Separate regressions for the standards of proof reveal that the effect is only present in the preponderance of evidence condition (beta=.21, p=.06) but not in the beyond reasonable doubt condition (beta=.07, p=.22).

Finally, it is revealing to consider the interrelation between subjective probability of guilt and the level of confidence, per choice and per standard of proof. The result is graphic. Subjects only convict if their estimate of guilt is above 50%[7]. While some convict under beyond reasonable doubt at a disturbingly low subjective probability of guilt, the majority of red diamonds is in the normatively expected places: subjects require a higher subjective probability under the stricter standard. If they acquit under preponderance of the evidence, their subjective probability of guilt is not above 50%. If they acquit although the subjective probability is above 50 %, this is always under the beyond a reasonable doubt instruction. This is further evidence for our finding that the standard of proof manipulation works. Finally, confidence is as one would expect it: Subjects are very confident if subjective probability of guilt is high and they convict, or if subjective probability of guilt is low and they acquit.

Figure 6
Subjective Probability vs. Level of Confidence

6. Discussion

We have been able to show that (mock) jury members do indeed rely on intuition when it comes to drawing inferences from visibly incomplete and conflicting evidence. The existing PCS theory on the mental mechanism leading to the formation of intuitions invites a normatively troubling hypothesis: the difference between standards of proof might be muted. This concern is not sup-

7 The one outlier notwithstanding.

ported by our data. In our experiments, the standard of proof manipulation has the normatively desired effect. If instructed to convict only if guilt is beyond a reasonable doubt, subjects are much more cautious. While coherence shifts are indeed stronger under the beyond a reasonable doubt instruction, this result is not driven by an excessive devaluation of exculpating evidence. On the contrary, under this instruction participants devalue inculpating evidence less strongly if they acquit. This is exactly what the legal order desires. Yet in their translations of standards into probabilities required for conviction, our subjects discriminate less than normatively desired. Their (explicit) subjective norms are not strict enough under beyond a reasonable doubt, and their norms are too strict under preponderance of the evidence. However conviction rates indicated that their implicit norms under both standards were even stricter than normatively expected; conviction rates were generally low. Finally, participants do not pay due regard to explicit information about probabilities.

The external validity of data using mock juries has been questioned (Bornstein and McCabe 2005; Cahoy and Ding 2006; Breau and Brook 2007). Yet given our research question, these limitations should not matter greatly. The real-life analogue of our prime dependent variable is first votes (Garvey, Hannaford-Agor et al. 2004), so that later interaction among jury members does not matter. Moreover, the experiment gives us full control over independent variables, which could not be achieved otherwise. We acknowledge another limitation of our findings: we are not addressing the later phase of deliberation (on this see Devine, Buddenbaum et al. 2007).

Our results do not call for institutional intervention for the sake of defendants in criminal procedure. In the normatively most critical situation, the instructions do a reliable job. We acknowledge that our subjects do not interpret standards of proof in the normative way. Yet the overly lenient interpretation of beyond a reasonable doubt does not translate into too many convictions. If alpha errors, i.e. undue convictions, are what the legal order is mainly concerned about, there does not seem to be reason for readjustment. Beta errors, i.e. unwanted acquittals, are a different matter. Moreover, our setting was not tuned to test subjects on tasks where intuition is particularly likely to go astray, like tasks inviting anchoring (Strack and Mussweiler 1997) or a hindsight bias (Fischhoff 1975). Depending on context, the legal order might therefore be tempted to intervene into the mental mechanism of jury members (Guthrie, Rachlinski et al. 2007).

Any intervention aimed at bringing outcomes closer to the norm should, however, keep three concerns in mind. While experimenters sometimes are in a position to reduce the beta error without increasing the alpha error, e.g. by increasing sample size, it is much harder to do that in court. Chances are that the intervention also makes false convictions more likely. At least in criminal procedure this is highly undesirable. It is probable that jurors would become better at assessing explicit information on conditional probabilities if they are given a chance to gain experience on the respective task. Yet this would run counter to the very idea of jury decision making. Jurors should neither be selected for expertise, nor should they gain experience on the job. Finally, interventions stressing the proper handling of explicit probabilities in a deliberate manner (as suggested by Guthrie, Rachlinski et al. 2007) would curb the power of intuition. This

would deprive jurors of the most powerful mental tool for adequately treating those manifestly ill defined problems of inference that are typical for judicial procedure.

Our approach and our results point the way for a directed improvement which circumvents such negative side-effects. Information on probabilities could be presented in a more accessible way. And supplementary evidence could be introduced that gives jury members a chance to put the evidence at hand in perspective. For the former purpose, probabilistic information could be presented graphically or in a frequency format, which both has been shown to enhance comprehension (Sloman, Over et al. 2003) For the latter purpose, additional evidence could be made admissible that gives jury members a sense how strong the evidence tends to be in like cases. Given the available knowledge about mental mechanism, the legal order should become attentive to the preconditions for constructing proper mental representations of the case.

References

Bar-Hillel, M. (1980). "The Base-Rate Fallacy in Probability Judgments." Acta Psychologica **44**: 211-233.

Barbey, A. K. and S. A. Sloman (2007). "Base-rate Respect. From Ecological Rationality to Dual Processes." Behavioral and Brain Sciences **30**: 241-254.

Bornstein, B. H. and S. G. McCabe (2005). "Jurors of the Absurd? The Role of Consequentiality in Jury Simulation Research." Florida State University Law Review **32**: 443-467.

Bowers, W. J., W. D. Foglia, et al. (2006). "The Decision Maker Matters. An Empirical Examination of the Way the Role of the Judge and the Jury Influence Death Penalty Decision-Making." Washington Lee Law Review **63**: 932-1010.

Breau, D. L. and B. Brook (2007). ""Mock" Mock Juries. A Field Experiment on the Ecological Validity of Jury Simulations." Law and Psychology Review **31**: 77-92.

Bright, D. A. and J. Goodman-Delahunty (2006). "Gruesome Evidence and Emotion: Anger, Blame, and Jury Decision-Making." Law and Human Behavior **30**: 183-202.

Cahoy, D. R. and M. Ding (2006). "The Stakes Matter. Empirical Evidence of Hypothetical Bias in Case Evaluation and the Curative Power of Economic Incentives." St. John's Law Review **80**: 1275-1305.

Clermont, K. M. and E. Sherwin (2002). "A Comparative View of Standards of Proof." American Journal of Comparative Law **50**: 243-276.

Day, J. G. (1987). "How Judges Think. Verification of the Judicial Hunch." Journal of Contemporary Legal Issues **1**: 73-105.

Devine, D. J., J. Buddenbaum, et al. (2007). "Deliberation Quality. A Preliminary Examination in Criminal Juries." Journal of Empirical Legal Studies **4**: 273-303.

Dreyfus, H. L., S. E. Dreyfus, et al. (1986). Mind over Machine. The Power of Human Intuition and Expertise in the Era of the Computer. New York, Free Press.

Eisenberg, T., P. L. Hannaford-Agor, et al. (2005). "Judge-Jury Agreement in Criminal Cases: A Partial Replication of Kalven and Zeisel's The American Jury." Journal of Empirical Legal Studies **2**: 171-206.

Eisenberg, T., P. L. Hannaford-Agor, et al. (2006). "Juries, Judges, and Punitive Damages. Empirical Analyses Using the Civil Justice Survey of State Courts 1992, 1996, and 2001 Data." Journal of Empirical Legal Studies **3**: 263-295.

Eisenberg, T. and G. P. Miller (2007). "Do Juries Add Value? Evidence from an Empirical Study of Jury Trial Waiver Clauses in Large Corporate Contracts." Journal of Empirical Legal Studies **4**: 539-588.

Engel, C. (2008). Institutions for Intuitive Man. Better Than Conscious? C. Engel and W. Singer. Cambridge, MIT Press**:** 391-410.

Finucane, M., A. Alhakami, et al. (2000). "The Affect Heuristic in Judgements of Risks and Benefits." Journal of Behavioral Decision Making **13**: 1-17.

Fischhoff, B. (1975). "Hindsight Is Not Equal to Foresight. The Effect of Outcome Knowledge on Judgment under Uncertainty." Journal of Experimental Psychology: Human Perception & Performance **1**: 288-299.

Franklin, J. (2006). "Case Comment - United States v. Copeland, 369 F.Supp.2d 275 (E.D.N.Y. 2005): Quantification of the 'Proof Beyond Reasonable Doubt' Standard." Law, Probability and Risk **5**: 159-165.

Garvey, S. P., P. L. Hannaford-Agor, et al. (2004). "Juror First Votes in Criminal Trials." Journal of Empirical Legal Studies **1**: 371-398.

Gigerenzer, G., P. M. Todd, et al. (1999). Simple Heuristics that Make us Smart. New York, Oxford University Press.

Givelber, D. and A. Farrell (2008). "Judges and Juries: The Defense Case and Differences in Acquittal Rates." Law and Social Inquiry **33**: 31-52.

Glöckner, A. (2008). How Evolution Outwits Bounded Rationality. The Efficient Interaction of Automatic and Deliberate Processes in Decision Making and Implications for Institutions. Better Than Conscious? C. Engel and W. Singer. Boston, MIT Press**:** 259-284.

Glöckner, A. and T. Betsch (2008). "Modeling Option and Strategy Choices with Connectionist Networks. Towards an Integrative Model of Automatic and Deliberate Decision Making." Judgement and Decision Making **3**: 215-228.

Glöckner, A. and T. Betsch (2008). "Multiple-Reason Decision Making Based on Automatic Processing." Journal of Experimental Psychology: Learning, Memory and Cognition ***: ***.

Guthrie, C., J. J. Rachlinski, et al. (2007). "Blinking on the Bench. How Judges Decide Cases." Cornell Law Review **93**: 1-43.

Hastie, R., S. Penrod, et al. (1983). Inside the Jury. Cambridge, Mass., Harvard University Press.

Heller, K. J. (2006). "The Cognitive Psychology of Circumstancial Evidence." Michigan Law Review **105**: 241-305.

Holyoak, K. J. and D. Simon (1999). "Bidirectional Reasoning in Decision Making by Constraint Satisfaction." Journal of Experimental Psychology: General **128**: 1-29.

Hsee, C. K. (1996). "The Evaluability Hypothesis: An Explanation for Preference Reversals between Joint and Separate Evaluations of Alternatives." Organizational Behavior and Human Decision Processes **67**: 247-257.

Jackson, J. D. (1996). "Analysing the New Evidence Scholarship: Towards a New Conception of the Law of Evidence." Oxford Journal of Legal Studies **16**: 309-328.

Johnson-Laird, P. N. (1983). Mental Models. Towards a Cognitive Science of Language, Inference and Consciousness. Cambridge, Cambridge University Press.

Kagehiro, D. K. (1990). "Defining the Standard of Proof in Jury Instructions." Psychological Science **1**: 194-200.

Kagehiro, D. K. and W. C. Stanton (1985). "Legal vs. Quantified Definitions of Standard of Proof." Law and Human Behavior **9**: 159-178.

Kahneman, D. and A. Tversky (1972). "Subjective Probability. A Judgement of Representativeness." Cognitive Psychology **3**: 430-454.

Kaye, D. (1982). "The Limits of the Preponderance of the Evidence Standard. Justifiable Naked Statistical Evidence and Multiple Causation." Law and Social Inquiry **7**: 487-516.

King, N. J. and R. L. Noble (2005). "Jury Sentencing in Noncapital Cases. Comparing Severity and Variance with Judicial Sentences in Two States." Journal of Empirical Legal Studies **2**: 331-367.

Koehler, J. and W. C. Thompson (2006). "Mock Jurors' Reactions to Selective Presentation of Evidence from Multiple-Opportunity Searches." Law and Human Behavior **30**: 455-468.

Lillquist, E. (2002). "Recasting Reasonable Doubt. Decision Theory and the Virtues of Variability." UC Davis Law Review **36**: 85-198.

Markus, H. and R. B. Zajonc (1985). The Cognitive Perspective in Social Psychology. Handbook of Social Psychology. G. Lindzey. New York, Random House**:** 137-230.

Orloff, N. and J. Stedinger (1983). "A Framework for Evaluating the Preponderance of the Evidence Standard." University of Pennsylvania Law Review **131**: 1159-1174.

Pennington, N. and R. Hastie (1986). "Evidence Evaluation in Complex Decision Making." Journal of Personality and Social Psychology **51**: 242-258.

Pennington, N. and R. Hastie (1988). "Explanation-Based Decision Making. Effect of Memory Structure on Judgement." Journal of Experimental Psychology: Learning, Memory and Cognition **14**: 521-533.

Pennington, N. and R. Hastie (1991). "A Cognitive Theory of Juror Decision Making. The Story Model." Cardozo Law Review **13**: 519-557.

Pennington, N. and R. Hastie (1992). "Explaining the Evidence. Tests of the Story Model for Juror Decision Making." Journal of Personality and Social Psychology **62**: 189-206.

Pennington, N. and R. Hastie (1993). "Reasoning in Explanation-Based Decision-Making." Cognition **49**: 123-163.

Pennington, N. and R. Hastie (1993). The Story Model for Juror Decision Making. Inside the Juror. The Psychology of Juror Decision Making. R. Hastie. Cambridge, Cambridge University Press**:** 192-221.

Pettys, T. E. (2007). "The Emotional Juror." Fordham Law Review **76**: 1609-1640.

Read, J. D., D. A. Connolly, et al. (2006). "An Archival Analysis of Actual Cases of Historic Child Sexual Abuse: A Comparison of Jury and Bench Trials." Law and Human Behavior **30**: 259-285.

Robbennolt, J. K. (2005). "Evaluating Juries by Comparison to Judges. A Benchmark for Judging?" Florida State University Law Review **32**: 469-509.

Saunders, H. D. (2005). Quantifying Reasonable Doubt: A Proposed Solution to an Equal Protection Problem.

Schmechel, R. S., T. P. O'Toole, et al. (2006). "Beyond the Ken? Testing Jurors' Understanding of Eyewitness Reliability Evidence." Jurimetrics **46**: 177-214.

Simon, D. (2004). "A Third View of the Black Box. Cognitive Coherence in Legal Decision Making." University of Chicago Law Review **71**: 511-586.

Simon, D. and K. J. Holyoak (2002). "Structural Dynamics of Cognition. From Consistency Theories to Constraint Satisfaction." Journal of Personality and Social Psychology **6**: 283-294.

Simon, D., D. C. Krawczyk, et al. (2008). "The Transience of Constructed Preferences." Journal of Behavioral Decision Making **21**: 1-14.

Simon, D., D. C. Krawczyk, et al. (2004). "Construction of Preferences by Constraint Satisfaction." Psychological Science **15**: 331-336.

Simon, D., L. B. Pham, et al. (2001). "The Emergence of Coherence Over the Course of Decision Making." Journal of Experimental Psychology: Learning, Memory and Cognition **27**: 1250-1260.

Simon, D., C. J. Snow, et al. (2004). "The Redux of Cognitive Consistency Theories: Evidence Judgments by Constraint Satisfaction." Journal of Personality and Social Psychology **86**: 814-837.

Sloman, S. A., D. Over, et al. (2003). "Frequency illusions and other fallacies." Organizational Behavior and Human Decision Processes **91**(2): 296-309.

Slovic, P., M. Finucane, et al. (2002). The Affect Heuristic. Heuristics and Biases. The Psychology of Intuitive Judgement. T. Gilovich, D. W. Griffin and P. Slovic. New York, Cambridge University Press**:** 397-420.

Spencer, B. D. (2007). "Estimating the Accuracy of Jury Verdicts." Journal of Empirical Legal Studies **4**: 305-329.

Steblay, N., H. M. Hosch, et al. (2006). "The Impact on Juror Verdicts of Judicial Instruction to Disregard Inadmissible Evidence: A Meta-Analysis." Law and Human Behavior **30**: 469-492.

Strack, F. and T. Mussweiler (1997). "Explaining the Enigmatic Anchoring Effect. Mechanisms of Selective Accessibility." Journal of Personality and Social Psychology **73**: 437-446.

Taruffo, M. (2003). "Rethinking the Standards of Proof." American Journal of Comparative Law **51**: 659-677.

Tillers, P. and J. Gottfried (2006). "Case comment—United States v. Copeland, 369 F. Supp. 2d 275 (E.D.N.Y. 2005): A Collateral Attack on the Legal Maxim That Proof Beyond A Reasonable Doubt Is Unquantifiable?" Law, Probability and Risk **5**: 135-157.

Weinstein, J. B. and I. Dewsbury (2006). "Comment on the Meaning of 'Proof Beyond a Reasonable Doubt'." Law, Probability and Risk **5**: 167-173.

Appendix: Case Description (medium probability)

Background: Hans H.

Hans H. is 34 years old. He lives in Frankfurt/Main with his wife, Katrin, and two children. Hans works for the large construction firm "Hausbau GmbH" (Hausbau Ltd.). After having worked as a foreman for more than two years, he complained to his superior that the job caused him back trouble. His boss then offered Hans a job in the company's administration offices, assigning him the role of construction manager. Hans' task was to supervise the progress made on the various building projects and to coordinate the different groups. Hans is generally considered to be a hard-working colleague. His colleagues say that he often seems reserved and at times even a little grumpy.

At the end of each day, the company's accountant places all company cash in the safe. This safe is located at the rear of the accounts office. The safe is also used to store other sensitive documents, including bids and project reports.

Apart from the accountant and her assistant, the construction managers, sales managers and managers have access to the safe. All in all, 18 people, including Hans, can use the safe. The safe has a time mechanism that records when the safe is opened and closed. One morning the accountant noticed that € 5,200 in cash was missing. The time mechanism showed that the safe had last been opened on the previous evening at 7:14 pm. After an investigation by a private detective, the firm instituted criminal proceedings against Hans H.

You will now be presented with the evidence from both parties. All witnesses have sworn under oath to make statements that correspond to the truth only and have been warned that false statements can lead to criminal proceedings for perjury.

Please read the evidence carefully and try to understand everything. Take as much time to do this as you deem necessary. You do not have to learn the evidence off by heart – whenever necessary, you will be able to consult it again.

Synopsis of the Evidence

A CCTV camera, installed at the entrance of the office building, shows a car rapidly leaving a parking space in front of the building at 19:17 pm on the evening in question. However, the picture was out of focus and the detective was unable to read the license plate. The video shows a white XY car. The make of Hans H.'s car is XY, it is white, and he seen was seen driving to work in it on the day in question. According to the detective, 6% of all cars in the area are white XY cars. He also found out that Hans paid back a loan of € 4,870 to his bank one day after the money had disappeared. The debts had accumulated in the last three months, and the bank had already threatened to take legal action. Hans testified that he had taken out the loan to help his sister-in-law, who runs a flower shop in Aachen. She gave him back the money in cash and he used it to pay back the loan. Hans explained that he cannot prove this cash transfer with receipts, since in the floral business larger financial transactions are sometimes conducted in cash.

Silvia, a manager of "Hausbau GmbH", testified that she saw Hans at 8 pm on the evening in question, when they both picked up their children from an event at the school. Hans was wearing elegant trousers and a jacket he had not worn at work. Silvia testified that it takes between 45 and 50 minutes at that time of day to get from the office to the school at the other end of town.

Hans testified that he has not had a criminal record for the last 16 years. At the age of 18, he was arrested for attempting to break into an apartment. He was convicted of this offence. Since then, he has never again been in conflict with the law.

A few months before the incident, Hans had been summoned by his boss to discuss the payment of expenses claimed by Hans. Visibly annoyed, the boss had given out to Hans for claiming certain expenses with no justification. Hans had argued that other construction managers had been claiming the same expenses and that the boss had therefore been challenging him unjustly. His boss had disagreed, refusing to reimburse these costs and also making clear to him that a promotion he had already been promised would fall through on account of these events. Hans had been deeply hurt by this. In the following weeks, he had quite frequently been seen working late at the office.

A technician who had been called to repair the photocopier testified that he had seen someone leave the accounts office in great haste at about 7.15 pm. When questioned by the detective a day after the incident, the technician identified this person as Hans. When asked how sure he was about this, the technician said he was "at least 80%" certain. He explained that he had seen Hans once or twice before in the office.

The Arguments Presented by the Parties

You will now be given the arguments presented by the company's lawyer, and then those presented by Hans' lawyer. In the light of these arguments, you will later be asked to evaluate the case.

Arguments Made by the Company

- The fact that only 6% of cars in the area are white XY cars makes it very likely that it was Hans who was filmed leaving the parking lot.
- It is no coincidence that Hans paid back his loan exactly one day after the burglary. He paid off his debts with the money he had stolen from the company safe.
- It is unlikely that larger financial transactions are conducted in cash in the floral business.
- Hans could have hurried up in order to be at the school for 8 pm.
- No matter how heavy the traffic, if one drives aggressively enough it is possible to shorten the journey time by a significant margin.
- In general, it is very likely that people who have already committed a crime will do so again at a later stage.
- Hans was annoyed by the sanctions imposed on him by his boss. Stealing the money from the safe was a possibility to take revenge on the company.
- In general, one can assume that people who feel unjustly treated do mean things.
- The fact that the technician was at least 80% certain in his identifying Hans as the man who left the accounts office proves that Hans stole the money.
- One can generally assume that people correctly identify other people, particularly when they have seen them before.

Arguments Made by the Defence

- The fact that a high 6% of cars in the area are white XY cars makes it less likely that it was Hans who was filmed leaving the parking lot.
- Hans paid back his debts with the money he received from his sister-in-law.
- In the floral business, larger financial transactions are indeed sometimes conducted in cash.
- It was virtually impossible for Hans to drive from the office to the school and be there for 8 pm, changing his clothes on the way.
- In evening rush hour traffic, it is very difficult to shorten the journey time, even if one drives as aggressively as possible.
- It is wrong to assume that people who have committed a crime will commit another.
- Hans did not want to take revenge on the company for his unfair treatment; instead, he tried to work even harder in order to prove himself to his boss.
- In general, one can assume that people who feel unjustly criticized in their job tend to work harder in order to prove themselves.
- The fact that the technician was not certain in his identification of Hans means that the person who took the money could have been someone else.
- One can generally assume that people often make mistakes when identifying other people, if they have only seen them once or twice before.