J.W. de Keijser & P.J. van Koppen (2007). Paradoxes of proof and punishment: Psychological pitfalls in judicial decision making. *Legal and Criminological Psychology*, *12*, 189-205.

Legal and Criminological Psychology (2007), 12, 189–205

The British Psychological Society

www.bpsjournals.co.uk

Paradoxes of proof and punishment: Psychological pitfalls in judicial decision making

Jan W. de Keijser and Peter J. van Koppen*

Netherlands Institute for the Study of Crime and Law Enforcement (NSCR), Leiden, The Netherlands

Purpose. This study focuses on two psychological mechanisms that may inadvertently affect judges' decisions on proof of guilt and on punishment. It involves mechanisms that are clearly in conflict with formal judicial doctrine. One hypothesis, the *conviction paradox*, asserts that, faced with very serious offences, a judge's standard of proof will be lower than for less serious, but otherwise comparable, offences. A second hypothesis, *compensatory punishment*, asserts that in cases with relatively weak evidence, judges who nevertheless render a guilty verdict will be inclined to compensate their initial doubt on the matter of guilt by meting out a less severe sentence.

Method. The hypotheses are evaluated in an experiment with Dutch judges and justices who serve in criminal courts. This was done using fictitious but highly realistic dossiers of criminal cases.

Results. Neither of the two hypotheses was supported in the present study.

Conclusions. Findings are discussed in relation to their implications for theory development and future research in the area of legal decision making.

In this study we examine two particular psychological mechanisms that may inadvertently affect judges' decisions on guilt and on punishment using an experiment with members of the Dutch criminal court judiciary. One mechanism, the conviction paradox, involves differential weighing of evidence and subsequent decisions of guilt. The other, compensatory punishment, explicitly connects the degree of certainty that preceded a guilty verdict to the severity of sentence (following a proposal made by Wagenaar, van Koppen, & Crombag, 1993).

Depending on the national justice system (and often also on the type of adjudicated case), decisions on the matter of guilt in criminal cases are taken either by judges or by juries (Damaška, 1986; Kritzer, 2002). While all Western legal systems have detailed rules governing the admissibility of evidence, the decision on guilt remains one that can hardly be directed by strict formal criteria (Hawkins, 1986; Hutton, 1995;

*Correspondence should be addressed to Peter J. van Koppen, Netherlands Institute for the Study of Crime and Law Enforcement (NSCR), P.O. Box 792, 2300 AT Leiden, The Netherlands (e-mail: PvanKoppen@nscr.nl).

DOI:10.1348/135532506X118721

Lovegrove, 1999; Twining, 1991, 1995; Wagenaar *et al.*, 1993). The same holds for the sentencing decision. In fact, 'the social practice of selecting the appropriate type and amount of sentence has remained for the most part a substantive irrational process' (Hutton, 1995, p. 556).

Although in most systems 'beyond reasonable doubt' is the formal criterion for a guilty verdict, it is one that is impossible to objectify (Horowitz, 1997; Horowitz & Kirkpatrick, 1996; Sheppard, 2003; Solan, 1999; Stoffelmayr & Diamond, 2000; Sundby, 1989). By far, most empirical studies were on jury decision making, most of which used mock jurors (see for overviews Greene *et al.*, 2002; Saks, 1997). Studies demonstrated that a myriad of personality variables and social and psychological factors influence decisions on guilt.

Sentencing research is the other main area of empirical research on judicial decision making. Ever since the groundbreaking work by Hogarth (1971) and Hood (1972), disparity in sentencing and the influence of extra-judicial, social and psychological factors on sentencing has time and time again been demonstrated (Fitzmaurice & Pease, 1986; Lovegrove, 1999). While the areas of research on decisions on guilt and on sentencing have developed largely in isolation from one another, their focus on unwarranted extra-legal considerations in criminal justice decision making inextricably connects them.

In practice, a gap remains in criminal law decision making that cannot be filled by legal rules, despite legislators' efforts to curtail the human factor. Sentencing guidelines, for instance, can be transformed or even circumvented in practice (Engen, Gainey, Crutchfield, & Weis, 2003; Johnson, 2005; Kempf-Leonard & Sample, 2001; Ulmer & Kramer, 1998). The bulk of guidelines developed in Western criminal justice systems is voluntary or presumptive in character. Judges and juries retain discretion in verdicts and sentencing (Tonry & Hatlestad, 1997). Thus, the weighing of evidence, on the one hand, and the magnitude of the sentence, on the other hand, ultimately cannot be but subjectively decided by the fact finder, be it jury or judge.

Human decision making in criminal cases

In psychological research on decision making, a model has long been used in which decision makers were assumed to first weigh all relevant information in order to give a considered decision (Wald, 1947; Wason, 1966); a model resembling a statistical procedure (see for reviews Gilhooly, 1988; Oakhill & Garnham, 1993). In subsequent research it was shown that decision-making behaviour of individuals does not comply to that 'ideal' model and that humans do divert from it, depending on the circumstances (see for reviews Johnson-Laird, 1999; Manktelow, 1999; Shafir & LeBoeuf, 2002).

Although two decades of research have emphasized the shortcomings of human judgment and decision-making processes (Kleinmuntz & Schkade, 1993), there has been serious dissent from this position (Edwards, 1992; Hammond, 2000). The 'ideal' model implicitly assumes that decision makers have virtually unlimited time and means to assemble and weigh all relevant information. Research by Gigerenzer and colleagues demonstrated that decision makers usually employ so-called fast and frugal heuristics that, depending on the circumstances, can even generate better decisions than the 'ideal' model (Gigerenzer, 2002; Gigerenzer, Todd, & ABC Research Group, 1999; Todd & Gigerenzer, 2000). Fast and frugal models are simple process models that do not search through all available information, do not integrate all relevant information and base the decision on only one cue (Dhami & Harries, 2001; Gigerenzer & Todd, 1999). In his recent

Nobel prize acceptance speech, Kahneman (2003) further argued that most decision making is in fact quite intuitive and usually rather unproblematic and successful (he drew this from Klein, 1998).

The conviction paradox

The psychology of the decision maker may nevertheless have unwarranted influences in the decisions. We identify two psychological pitfalls in criminal justice decision making that have, as yet, not been subjected to systematic empirical research but may play a role in legal decision making. The first psychological pitfall we studied is the conviction paradox. This is based on signal detection theory.

Signal detection theory demonstrated how consequences of decision outcomes influence decision making (Banks, 1970; Egan, 1975; Green & Swets, 1966; Hirshman, Lanning, Master, & Henzler, 2002; Lockhart & Murdock, 1970; McNicol, 1972; Poor, 1994). The basis of this theory is the existence of two kinds of stimuli: a signal and a non-signal. The decision maker has to distinguish between these two, but is often handicapped by two things: a background of noise, and the fact that most signals are continuous, rather than discrete. In deciding, the decision maker can make two types of error: a false positive, i.e. accepting a non-signal as a signal, and a false negative, i.e. accepting a signal as a non-signal.

In judicial decision making, a guilty accused can be seen as a signal; an innocent accused as a non-signal. However, there are no discrete signals in judicial decision making: some defendants appear somewhat guilty, others somewhat more and still other defendants appear to be extremely guilty. The fact finder has to convert this continuous signal into a discrete decision: guilty or not. This can only be done by accepting a decision criterion and one should only convict if the 'level' of evidence surpasses this criterion and acquit in the other cases.

Signals in terms of strength of evidence from guilty and innocent defendants do overlap. Thus, choosing the location of the decision criterion always involves a trade off. A higher level of required certainty of proof causes more innocent, but also more guilty defendants to be acquitted. On the other hand, a lower level causes the reverse.

Determining the location of the decision criterion also involves a trade off between the risk of a false positive against the risk of a false negative. Legal doctrine prescribes that the costs of false positives outweigh the costs of false negatives (see for instance Volokh, 1997; Wigmore, 1970). That may be illustrated by the oft heard judicial maxim: 'Better to acquit ten guilty persons than to convict a single innocent' (Volokh, 1997), regardless of the seriousness of the crime.

This jurisprudential maxim is not always reflected in jurors' decision making (compare, for instance, Dane, 1985; Hastie, 1993; Hastie, Penrod, & Pennington, 1983; Horowitz & Kirkpatrick, 1996; Kerr *et al.*, 1976; MacCoun & Kerr, 1988; Nagel, Lamm, & Neef, 1981; Simon & Mahan, 1971). There a pattern emerges where jurors adopt a higher standard of reasonable doubt when the crime is more serious (Horowitz, 1997; Horowitz & Kirkpatrick, 1996; Kagehiro & Stanton, 1985; Kerr, 1978; Simon, 1969; Simon & Mahan, 1971; Stoffelmayr & Diamond, 2000). Please note that jurors in addition display a leniency bias. In studies where the evidence is balanced carefully (as is prescribed to find any relatively subtle effect, following Devine, Clayton, Dunford, Seying, & Pryce, 2001; Saks, 1997), juries show a tendency to acquit following the reasonable doubt standard of proof, because that standard favours the defendant (Davis, 1980; Koch & Devine, 1999; MacCoun & Kerr, 1988; Stasser, Kerr, & Bray, 1982;

Tindale, Davis, Vollrath, Nagao, & Hinsz, 1990). None of these studies, however, compared decision making in serious and less serious cases.

Signal detection theory, however, would predict an inverse relation between crime seriousness and the location of the decision criterion. One important utility that influences the position of the criterion is formed by the consequences that false positive and false negative decisions may have. Such potential outcomes induce different costs (see Forst, 2004). The costs of a false positive decision (i.e. wrongful conviction) is unjustified suffering of the person convicted. The costs of a false negative decision is that a person who is in fact guilty of a crime but is nevertheless acquitted, continues to pose a danger to society; a danger that may very well materialize in continued re-offending. These costs are small if it concerns, for instance, a shoplifter, whereas the costs are quite high if the case concerns an accused who is prosecuted for serial rape. In this line of reasoning, the more serious the crime and the more dangerous the offender, the higher the potential costs are of a false negative decision. For the decision maker confronted with a serious case, the costs of a false negative decision may outweigh the costs of a false positive decision. This produces a paradox. While one might expect fact finders to be especially careful when considering the evidence in more serious cases, they may rather be inclined to satisfy themselves with a lower degree of certainty because the costs of a false negative in such cases are considered tremendously high. (There are more reasons why it can be expected that for very serious offences fact finders accept a lower level of proof. These are discussed by Gross, 1996.)

It could be argued that as a consequence of both a false negative and a false positive decision, the real offender remains at large. This means that a false positive decision would always bear more costs than a false negative decision. In both, the real perpetrator may continue offending, while after a false positive decision in addition the wrongly accused suffers a sentence. From a general perspective, this may be true. From the perspective of the fact finder, however, this is only true if the fact finder has any influence on what is happening to the real perpetrator when confronted with an accused who in reality is innocent. Fact finders cannot exercise such influence. After a case has been sent to the prosecution, the police almost never undertake any further action, even if the accused is acquitted and even if the court makes clear that in its opinion the accused is not acquitted just because of lack of evidence, but because the court is convinced he did not commit the crime. Thus, if the wrong individual is prosecuted, almost by definition the real offender remains at large. From the perspective of the fact finder, therefore, a false negative decision in which the real offender is released is the only one that may have grave societal consequences as a direct result of his decision.

Compensatory punishment

In recent years, psychological research has demonstrated that anticipated regret for possible outcomes of decisions plays a role in decision making (Connolly & Zeelenberg, 2002; Zeelenberg, 1999; Zeelenberg, van den Bos, van Dijk, & Pieters, 2002). It has been shown that when confronted with a choice between different alternatives, anticipated regret for each of the alternatives is weighed, generally resulting in the 'safe option' being preferred (Zeelenberg, 1999). Regret aversion can, under some circumstances, have a major impact on decision making. While this sometimes leads to inaction, it generally results in choosing the least risky option in terms of anticipated regret (Kahneman & Tversky, 1982; Seta, McElroy, & Seta, 2001; Zeelenberg *et al.*, 2002).

The judge is in a unique position from a 'regret perspective'. In the Netherlands and in other countries where a judge may sit without a jury, a judge makes consecutive decisions pertaining to the same criminal case: first on guilt and then (if found guilty) on the sentence. We can thereby extend the study of regret in decision making to the role of anticipated regret in consecutive decisions.

Suppose a judge renders a guilty verdict. Before he formally decided that the defendant is guilty, however, the judge did experience some hesitation on the issue of guilt. This enhances the probability that the judge anticipates regret (see Pieters & Zeelenberg, 2002, cited in Connolly & Zeelenberg, 2002). Against that backdrop, the judge now has a sentencing decision to make. Given the consecutive nature of this decision-making procedure, the judge has the unique cognitive opportunity to compensate the anticipated regret by giving a relatively lenient sentence. This is what we call compensatory punishment.

There is some support for compensatory punishment in research in legal decision making. Davis and colleagues (Davis, Holt, Spitzer, & Stasser, 1981, p. 12) discuss Gleisser (1968) who reported that French juries in the nineteenth century were reluctant to convict defendants because in those days judges usually imposed excessive sentences (see also Mello & Robson, 1985). They hypothesized that juries who also control sentencing might be induced to a greater willingness to convict. Their hypothesis was not supported with a significant effect. In civil cases, however, a comparable effect was found (Wissler, Kuehn, & Saks, 2000). Some studies show that juries decrease awards if their assessments of the plaintiff's negligence increased (Hammitt, Carroll, & Relles, 1985; Horowitz & Bordens, 1988; Shanley, 1985; Thomas & Parpal, 1987; Viscusi, 1988). As a consequence, awards in negligence cases are discounted twice, because subsequently the court further reduces the already lowered awards in proportion to the plaintiff's negligence (Hammitt *et al.*, 1985; Shanley, 1985). Other studies, however, did not find an influence on jury awards of contributory negligence by the plaintiff (Horowitz & Bordens, 1990; Vidmar, Lee, Cohen, & Stewart, 1994; Wissler, Evans, Hart, Morry, & Saks, 1997). We are unaware of empirical studies of this kind of compensation in criminal cases.

The conviction paradox and compensatory punishment

The conviction paradox is not a precondition for compensatory punishment. It may be evident from our perspectives, however, that the conviction paradox can very well serve as a catalyst for compensatory punishment. If a judge accepted a low level of proof in a serious case, that may induce anticipated regret which in turn may lead to a relatively low sentence.

These two psychological pitfalls have in common that they stand in contrast to judicial doctrine. Dutch formal doctrine rules them out, as doctrine does in most other Western countries. The conviction paradox involves different standards of reasonable doubt for different types of cases. There is no legal basis for such differences. Compensatory punishment circumvents something else. In legal systems with a two-stage procedure, where the jury decides on the guilt of the accused and the judge on the sentence, the two decisions are by nature separated (see, however, Hoffman, 2003). But also in non-jury cases in such systems and in legal systems where the court decides both on guilt and on the sentence, legal doctrine prescribes two separate decisions: after guilt has been established, the sentence should assume guilt as a fact and not incorporate doubts about the accused's guilt (compare Damaška, 1986, 1997, 1998).

We carried out our study within the Dutch legal system, where lay participation in decision making in criminal cases is unknown. We expect that the conviction paradox, which only concerns the decision on the guilt of the accused, is likewise relevant for decisions by juries. Compensatory punishment, which involves both the decision on guilt and the sentence, can be hypothesized to occur within a single decision maker, such as a Dutch bench court.

All cases in the Netherlands are tried by professional judges. Plea bargaining is unknown: all cases are tried in full. Also, all court decisions can be appealed to the appellate court – without leave to appeal – where the case is tried *de novo*. Serious cases are always tried by a court of three judges (see for descriptions in English on Dutch criminal procedure van Koppen, 2002; Nijboer, 1999; Taekema, 2004; Tak, 2003).

Dutch criminal procedure is dominated by written records. All officials involved – the police, the prosecution, the judge-commissioner (judge of instruction), the courts, but also the defence – produce written records that become part of the official case file, the dossier. Dossiers include all important sources of evidence and information. In court, interaction between judges, prosecutor, accused and counsel focuses on evaluation of the dossier. In general, the parties make little use of their right to summon witnesses or experts at trial (Nijboer, 1999).

Dutch judges enjoy wide discretionary powers in choosing type and severity of punishment (Tak, 1997). The penal code specifies minimum terms for punishments in general (e.g. 1-day imprisonment) and specific maximum terms are specified for each offence in the penal code. There are no sentencing guidelines, though Dutch judges do aim to enhance consistency through mutual consultation and by formulation of sentencing policies for clearly defined types of offences. Furthermore, the Dutch prosecutor requests a specific punishment at the end of the trial hearing. Judges are not bound by the requested punishment although it does provide some form of anchor point in judges' deliberations on the sentence.

Method

We provided judges with detailed case files closely resembling actual case dossiers that play such a dominant role in the Dutch legal procedure. We included in the dossiers all relevant information in the same raw format as judges are used to in reality, whilst still being able to systematically vary the factors relevant to our hypotheses. We thereby satisfy a number of objections – mainly related to ecological validity – to the use of experiments in legal decision making (cf. Koneçni & Ebbesen, 1992; Lovegrove, 1999).

> *Hypothesis 1.* Conviction paradox: In serious cases judges will sooner convict on the basis of relatively thin evidence than in less serious cases (*ceteris paribus*).
>
> *Hypothesis 2.* Compensatory punishment: An accused will receive a lesser punishment if the evidence against him is relatively thin than in the same case with strong evidence (*ceteris paribus*).

Materials

Two factors were manipulated: crime seriousness and evidence strength. Two versions of an assault case were constructed, resulting in (A) an aggravated assault case, and (B) a simple assault case. The only differences between the two versions (A and B) involved elements in the dossiers related to crime seriousness. This was manipulated by

varying the type and amount of violence applied by the perpetrator and subsequent injuries suffered by the victim. In the serious version (A), the offender did not only kick the body of the victim, but also his head, resulting in permanent loss of powers of speech as well as irreparable paralysis from the waist down. In the less serious version (B), only the body was kicked, not resulting in permanent injuries. Everything else was kept identical between the two cases. As a contrast to these two violent crimes, a third case was constructed: a burglary case (C).

Each of the three cases (aggravated assault, simple assault, burglary) were presented in two versions: one with strong and very convincing evidence, and one with legally sufficient, though relatively weak evidence. Thus, the material for the experiment consisted of six dossiers in total (see Table 1).

Creating strong evidence versions of the cases was rather straightforward. Two aspects were used to promote a conviction of guilt. In all three strong evidence versions, the accused confessed to having committed the crime, confirmed the story of witnesses and victims and was identified with a very high degree of certainty by witnesses.

Compared to the strong evidence cases, strength of evidence was scaled down in the weak evidence versions by having the accused deny and having witnesses identify him only after explicit hesitation. In addition, technical forensic evidence was indicative though inconclusive in the weak evidence versions. The lack of a real trial was compensated by a final sheet attached to the dossiers in which a short description was provided of what happened during the hypothetical trial. The sentence demanded by the prosecution was held constant over strong and weak evidence versions of the cases (see Table 1). These requested sentences were realistic ones consistent with national prosecution guidelines.[1]

In both strong and weak evidence versions, the same and substantial amount of circumstantial evidence against the accused was included, which was expected to start playing a role of importance in the weak evidence versions. All defendants lacked an alibi and had extensive criminal records including similar crimes committed in the past. The evidence in the weak versions of the dossiers was further tweaked after discussing the cases with professors of criminal law and criminal procedure as well as by having a number of experienced part-time judges evaluate the cases and afterwards discuss the strength of the evidence with us in detail.

Apart from a limited number of background questions, participants were requested to decide upon the matter of guilt of the accused and if pronounced guilty to give their sentence in free format. We did not use any scale with more measuring points for the verdicts, because that is alien to the criminal legal procedure. Judges were further asked to give reasons for the sentence (i.e. motivation), as judges are required to give under Dutch law.

Design

We used a complete between-subjects design. The conviction paradox was evaluated by comparing the number of guilty verdicts in the weak evidence versions of the serious and simple assault cases. We expected the simple assault case to generate fewer convictions. Compensatory punishment was evaluated with a comparison of average sentence length (given a guilty verdict) for the strong and the weak evidence versions. Thus, the design involves a 3 (case version) × 2 (strength of evidence) design.

[1] *The dossiers (in Dutch) are available from the authors.*

Table 1. The six dossiers and their content

	A Aggravated assault		B Simple assault		C Burglary	
Evidence	Strong	Weak	Strong	Weak	Strong	Weak
Total no. of pages	19	21	25	25	19	21
Total no. of words	6765	6991	8985	8562	6395	6680
Dossier elements						
Summary police findings	x	x	x	x	x	x
Indictment	x	x	x	x	x	x
Victim statement	o	o	x[1]	x[1]	x	x
Records of witness interviews	3	3	3	3	2	2
Record of technical forensic research	na	na	na	na	x	x
Report of arrest	x	x	x	x	x	x
Record of police interview with accused	x	x	x	x	x	x
Record of search in house accused	na	na	na	na	x	x
Record of photo-confrontation witness-accused	x	x	x	x	x	x
Results of forensic research	na	na	na	na	x	x
Record of second police interview with accused	o	x	o	x	x	x
Medical report on injuries victim	x	x	x	x	na	na
Psychological report on accused	x	x	o	o	o	o
Probation report on accused	o	o	x[2]	x[2]	x	x
Full criminal history of accused	x	x	x	x	x	x
Requisitoir: Summing up and punishment requested by prosecutor	x	x	x	x	x	x
Prison term requested by prosecution (months unsuspended)	30	30	2.5	2.5	6	6
Concise description of trial	x	x	x	x	x	x

[1] Victim states having no recollection of the incident (due to head injuries).
[2] Contents essentially the same as psychological report in case A.
x, included in dossier; o, not included in dossier; na, not applicable.

Procedure

We asked all 629 judges in district courts and justices in the court of appeals who serve in the criminal divisions of their courts to participate. We excluded the so-called replacement judges and justices, who serve part time in the courts, alongside their jobs elsewhere. The Dutch Council for the Administration of Justice (Raad voor de rechtspraak) wrote a letter of support to the presidents of all 19 district courts and five courts of appeal in the Netherlands, describing the study only in general terms as 'a study on legal decision making'. Two weeks later we sent the dossiers to the judges and justices (further denoted judges). A reminder was sent out 2 weeks later. Of course, participation was anonymous.

Response and representativeness

The overall response rate was 36.4% ($N = 229$). A limited number of background variables for the population were available to us, enabling a rough indication of representativeness of our sample: gender, type of judge (judge in court or justice in court of appeal) and regional dispersion grouped at the level of courts of appeal jurisdictions (see Table 2). Considering these variables, our sample may be considered quite representative of the population, although judges were slightly over-represented and justices slightly under-represented.

Table 2. Representativeness: gender, type of judge, and regional dispersion (percentages)

	All Dutch judges and justices	Our sample
N	629	229
Male	54	50
Female	46	50
Judge district court	80	86
Justice court of appeal	20	14
Regional dispersion		
Amsterdam	33	33
Arnhem	14	14
Den Haag	26	24
Den Bosch	18	19
Leeuwarden	9	9

Note. Regional dispersion grouped at the level of courts of appeal jurisdictions.

Results

Our experimental manipulation of the strength of the evidence was successful (see Table 3). In all three strong evidence cases, all judges rendered a guilty verdict. The numbers of acquittals in the weak evidence versions show that these dossiers provided ample opportunity for doubt. In motivating the sentence in the weak evidence versions, a number of judges explicitly noted that their guilty verdict was a close call. The margin for doubt, however, was, as intended, narrow enough to enable a majority of judges to decide upon a guilty verdict.

The decision of guilt is a simple discrete decision: yes or no. Nine participating judges, however, pronounced a guilty verdict for a different offence than the one specified in the indictment. These nine decisions were not included in further analyses.

Table 3. Decisions on guilt ($N = 220$; percentages)

		Guilty verdict		
Type of case	Evidence	Not guilty	Guilty	*N*
Aggravated assault	Strong	0	100	23
	Weak	23	77	51
Simple assault	Strong	0	100	27
	Weak	23	77	53
Burglary	Strong	0	100	22
	Weak	27	73	44

The judges who rendered a guilty verdict, subsequently gave a sentencing decision in free format. Coding these decisions was an easy task since, following a guilty verdict, all but four judges specified a prison sentence. The four exceptions involved community service orders. These were excluded from the analyses of the sentencing decisions.[2]

Decisions on guilt: The conviction paradox

If the conviction paradox affects judges' decision making, we would expect the percentage of guilty verdicts in the weak evidence version of the aggravated assault to be significantly higher than in the weak evidence version of the simple assault. Results show, however, no support for the hypothesis. In both cases, 77% of the judges decided upon a guilty verdict (see Table 3). Further analyses showed no significant effects of background variables on the decisions of guilt: there were no differences between male and female judges, between judges and justices, between experienced and relatively inexperienced judges and between regions.

Sentencing decisions

A prison sentence in the Netherlands can be imposed completely unsuspended, partly suspended or completely suspended. In all cases in this experiment, the large majority of judges specified a completely unsuspended prison term (74% in the aggravated assault case, 70% in the simple assault case, 83% in the burglary case). We have carried out our analyses on the total prison sentences. Analyses of only the unsuspended sentences did not produce different results. Three outliers were removed from further analysis. These involved relatively extreme sentences: one in the aggravated assault case (72 months), one in the simple assault case (8 months) and one in the burglary case (48 months). These outliers were at least 3 standard deviations from the respective means.

Compensatory punishment

If the judges are affected by compensatory punishment, we would expect judges in the weak evidence version of a case to sentence more leniently than judges in the strong evidence version of that same case.

[2] *Community service, if not conducted properly, can be converted to a prison sentence. In the Netherlands, a fixed scale is used to convert prison to community service and vice versa: 240 hours of community service equals 6 months in prison, while a lesser number of hours conforms to a lesser prison term pro rata. We also undertook our analyses with the four community service sentences converted to a prison term. This did not influence the results.*

The only case in which an effect in the predicted direction may be observed is the simple assault (see Table 4). This effect, however, is small (0.4 months) and not statistically significant. In the burglary case, the difference between the strong and weak evidence version is in the opposite direction and a mere (not statistically significant) 0.3 months. The most substantial difference between strong and weak evidence versions was in the aggravated assault case. The difference in this case is 4.2 months. However, it also constitutes an effect in the opposite direction as predicted and is not statistically significant.

Table 4. Sentences in weak and in strong evidence versions of the cases (convictions only; $N = 177$; months of imprisonment)

			Months of imprisonment		
Type of case	Evidence	*N*	Mean	*SD*	*t*-value
Aggravated assault	Strong	23	28.7	7.4	1.85, *ns*
	Weak	38	32.9	9.2	
Simple assault	Strong	23	3.1	1.2	−1.40, *ns*
	Weak	40	2.7	1.0	
Burglary	Strong	21	5.4	1.2	0.66, *ns*
	Weak	32	5.7	1.7	

Using the actual sample sizes and estimates of the standard deviations, we computed that the power of our tests is sufficient. For each of the cases concerned, a difference of a quarter in terms of average length of punishment issued in the strong evidence case would be detectable with a power of between 80% (simple assault) and 90% (both other cases) for a one-sided test with $\alpha = 5\%$.

Discussion

In our experiment, we did not find empirical support for the conviction paradox, nor for compensatory punishment. This may be considered surprising, since both mechanisms are derived from firmly established psychological phenomena that operate in many other fields of decision making. Our findings therefore invite further elaboration and explanation.

Two lines of reasoning may be adopted. In a first line of reasoning the position can be taken that the hypothesized mechanisms do occur in judges' decision making, despite findings from this experiment. One could suspect that compensatory punishment does not operate in cases in general, but only in incidental cases. If compensatory punishment is indeed a real mechanism but only in rare and incidental cases, it is not surprising that a more generalized experiment is unable to uncover it. The same argument could be made for the conviction paradox. Thus, the two phenomena may exist, may happen sometimes, be recognized by participants, are even open to qualitative analyses (Wagenaar *et al.*, 1993), but impossible to detect in a controlled experiment. Given this explanation, the fact remains that the case for the conviction paradox and compensatory punishment is now a weaker one than before we conducted our experiment.

A further argument in this first line of reasoning would be that our experiment lacked the finesse for a valid test of the hypotheses. This boils down to a methodological issue: true responsibility as felt by judges toward those directly and indirectly involved in real cases cannot be simulated in an experiment like ours. As such, one may object to our experiment that an important building block for the mechanisms of compensatory punishment and for the conviction paradox was missing, thus making it easier for participants to play safe, knowing that the cases were part of an empirical study. On the other hand, we believe that the experiment closely simulated the reality of Dutch criminal justice decision making. In practice, the work of Dutch judges is dominated by paper dossiers such as our experimental material. We cannot, however, deny that the experiment was unable to simulate the actual pressures that judges experience in real cases. This argument should, therefore, be kept in mind during our discussion.

In a second line of reasoning, we take the findings of the experiment at face value. Dutch judges and justices are invulnerable to the psychological pitfalls of both the conviction paradox and compensatory punishment. The question that now begs an answer is: Why? After all, both mechanisms were based on plausible and general psychological phenomena. Moreover, why would judges not be susceptible to the same basic psychological biases as are other people (Fitzmaurice & Pease, 1986)? If nothing else, it has become unlikely that these psychological pitfalls are common practice in Dutch criminal procedure. On the other hand, it seems unlikely that these expert judges were so different from many other experts who have failed to be invulnerable to biases. It could be expected that judges are even more vulnerable than many other types of experts, since they do not receive clear and fast feedback and lack explicit and well-defined theory as a basis for their judgments. It may be possible that this effect is countered by the judges using templates in decision making, thus making them less sensitive to particular biases in individual cases. It should be noted that the following explanation is (necessarily) mere conjecture. The answer to the question may be found in the nature of the Dutch judiciary. In the Netherlands, contrary to many other countries, any form of lay involvement in criminal justice decision making is absent. There are neither juries nor lay judges. The Dutch magistrature is exclusively professional with extensively trained and legally socialized judges. So much so that their professionalism, their *technicité*, may provide effective shielding against basic psychological pitfalls. Indeed, one of the main pillars in the training and socialization of magistrates is strict application of the formal hierarchical decision-making model discussed above.

In this line of reasoning, there are clear implications for further research. The 'professionalism explanation' can be tested in countries where lay fact finders participate. Lay fact finders, in general, lack the level of training and legal socialization that professional fact finders possess and, therefore, lack protection against particular psychological mechanisms. In some countries with lay participation in criminal procedure, typically jury systems such as in the United States and England and Wales, lay persons do not play a role in sentencing (an exception is the role of the American jury in capital cases; cf. Kritzer, 2002). In these countries, the conviction paradox could be tested using jurors, while compensatory punishment may operate among judges who, after the jury rendered a guilty verdict based on weak evidence, opt for a relatively low sentence (of course, such a study should exclude sentencing under mandatory sentencing guidelines). In some other countries, lay persons participate as lay assessors in a bench court together with one or more professional judges. Examples include Germany, Sweden, Switzerland and Poland. The case could be made that a mechanism

such as compensatory punishment shapes a 'haggling' process during deliberation between professional and lay judges. In such a haggling process, agreement to convict is reached in return for a relatively lenient sentence. Similar processes may underlie decision making in a mixed system. An example of such a mixed system is the jury in Belgium. In their *Hof van Assisen*, a jury of 12 decides on guilt of the accused, while subsequently the jury together with three professional justices of the court decide on the sentence.

A final word needs to be said on our experimental manipulation of evidence strength. The most telling difference between strong and weak evidence versions was whether or not the accused confessed. Although we are unaware of relevant studies, our work among the police, the prosecution and in the judiciary have taught us that not confessing tends to be regarded as not cooperating with the criminal procedure and frequently also as refusal to take responsibility for one's actions. Thereby, not confessing may raise the sentence, a fact that is often used by the police to induce a confession (cf. Gudjonsson, 2003; Vrij, 2003). A similar and more explicit mechanism may be observed in the USA, where defendants may be coerced to plead guilty by the threat of a relatively higher sentence following a jury trial (Sandefur, 2003). This mechanism thus could have counteracted compensatory punishment in our experiment. However, if this has indeed been the case, it may be interpreted as further evidence against compensatory punishment. Once it has been decided that the denying defendant is guilty, the judge will not be tempted to raise the sentence unless he or she strictly separated the decision of guilt from the sentencing decision.

In conclusion, our experiment raises doubts on the existence of the conviction paradox and of compensatory punishment among Dutch judges, at least to the extent that these are not common mechanisms. However, characteristics of Dutch criminal procedure in combination with the absence of lay participation have spawned a number of new research questions and opportunities for further study in an international comparative perspective.

Acknowledgements

We thank Cyrus Tata, Ron Huff, Michael Tonry and two anonymous reviewers for their comments on earlier drafts of this article. We thank Henk Elffers for support with the analyses.

References

Banks, W. P. (1970). Signal detection and human memory. *Psychological Bulletin*, *74*, 81–86.

Connolly, T., & Zeelenberg, M. (2002). Regret in decision making. *Current Directions in Psychological Science*, *11*, 212–216.

Damaška, M. R. (1986). *The faces of justice and state authority: A comparative approach to the legal process*. New Haven: Yale University Press.

Damaška, M. R. (1997). *Evidence law adrift*. New Haven: Yale University Press.

Damaška, M. R. (1998). Truth in adjudication. *Hastings Law Journal*, *49*, 289–308.

Dane, F. C. (1985). In search of reasonable doubt: A systematic examination of selected quantification approaches. *Law and Human Behavior*, *9*, 141–158.

Davis, J. H. (1980). Group decision and procedural justice. In M. L. Fishbein (Ed.), *Progress in social psychology* (Vol. 1, pp. 157–229). Hillsdale, NJ: Erlbaum.

Davis, J. H., Holt, R. W., Spitzer, C. E., & Stasser, G. (1981). The effects of consensus requirements and multiple decisions on mock juror verdict preferences. *Journal of Experimental Social Psychology*, *17*, 1–15.

Devine, D. J., Clayton, L. D., Dunford, B. B., Seying, R., & Pryce, J. (2001). Jury decision making: 45 years of empirical research on deliberating groups. *Psychology, Public Policy, and Law*, 7, 622–727.

Dhami, M. K., & Harries, C. (2001). Fast and frugal versus regression models of human judgement. *Thinking and Reasoning*, 7, 5–27.

Edwards, W. (1992). Discussion: Of human skills. *Organizational Behavior and Human Decision Processes*, *53*, 267–277.

Egan, J. P. (1975). *Signal detection theory and ROC analysis*. New York: Academic.

Engen, R. L., Gainey, R. R., Crutchfield, R. D., & Weis, J. G. (2003). Discretion and disparity under sentencing guidelines: The role of departures and structured sentencing alternatives. *Criminology*, *41*, 99–130.

Fitzmaurice, C., & Pease, K. (1986). *The psychology of judicial sentencing*. Manchester: Manchester University Press.

Forst, B. (2004). *Errors of justice: Nature, sources and remedies*. New York: Cambridge University Press.

Gigerenzer, G. (2002). *Adaptive thinking: Rationality in the real world*. New York: Oxford University Press.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer, P. M. Todd, & ABC Research Group, (Eds.), *Simple heuristics that make us smart* (pp. 3–34). New York: Oxford University Press.

Gigerenzer, G., Todd, P. M., & ABC Research Group (Eds.). (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Gilhooly, K. J. (1988). *Thinking: Directed, undirected and creative* (2nd ed.). New York: Academic.

Gleisser, M. (1968). *Juries and justice*. South Brunswick, NJ: A.S. Barnes.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Greene, E. L., Chopra, S. R., Kovera, M. B., Penrod, S. D., Rose, V. G., Schuller, R. A., & Studebaker, C. A. (2002). Jurors and juries: A review of the field. In J. R. P. Ogloff (Ed.), *Taking psychology and law into the twenty first century* (pp. 225–284). New York: Plenum.

Gross, S. R. (1996). The risks of death: Why erroneous convictions are common in capital cases. *Buffalo Law Review*, *44*, 469–500.

Gudjonsson, G. H. (2003). *The psychology of interrogations and confessions: A handbook*. Chichester: Wiley.

Hammitt, J. K., Carroll, S. J., & Relles, D. A. (1985). Tort standards and jury decisions. *Journal of Legal Studies*, *14*, 751–762.

Hammond, K. R. (2000). Coherence and correspondence theories in judgment and decision making. In T. Connolly, H. R. Arkes, & K. R. Hammond (Eds.), *Judgment and decision making: An interdisciplinary reader* (2nd ed., pp. 53–65). New York: Cambridge University Press.

Hastie, R. (1993). Algebraic models of juror decision processes. In R. Hastie (Ed.), *Inside the jury: The psychology of juror decision making* (pp. 84–115). Cambridge: Cambridge University Press.

Hastie, R., Penrod, S. D., & Pennington, N. (1983). What goes on in a jury deliberation. *American Bar Association Journal*, *69*, 1848–1853.

Hawkins, K. (1986). On legal decision-making. *Washington and Lee Law Review*, *43*, 1161–1242.

Hirshman, E., Lanning, K., Master, S., & Henzler, A. (2002). Signal-detection models as tools for interpreting judgements of recollections. *Applied Cognitive Psychology*, *16*, 151–156.

Hoffman, M. B. (2003). The case for jury sentencing. *Duke Law Journal*, *52*, 951–1010.

Hogarth, J. (1971). *Sentencing as a human process*. Toronto: University of Toronto Press.

Hood, R. (1972). *Sentencing the motor offender: A study of magistrates' views and practices*. London: Heinemann (Cambridge Studies in Criminology).

Horowitz, I. A. (1997). Reasonable doubt instructions: Commonsense justice and standard of proof. *Psychology, Public Policy, and Law*, *3*, 285–302.

Horowitz, I. A., & Bordens, K. S. (1988). The effects of outlier presence, plaintiff population size, and aggregation of plaintiffs on simulated civil jury decisions. *Law and Human Behavior*, *12*, 209–229.

Horowitz, I. A., & Bordens, K. S. (1990). An experimental investigation of procedural issues in complex tort trials. *Law and Human Behavior*, *14*, 269–285.

Horowitz, I. A., & Kirkpatrick, L. C. (1996). A concept in search of a definition: The effects of reasonable doubt instructions on certainty of guilt standards and jury verdicts. *Law and Human Behavior*, *20*, 655–670.

Hutton, N. (1995). Sentencing, rationality, and computer technology. *Journal of Law and Society*, *22*, 549–570.

Johnson, B. D. (2005). Contextual disparities in guidelines departures: Courtroom social contexts, guidelines compliance, and extralegal disparities in criminal sentencing. *Criminology*, *43*, 761–796.

Johnson-Laird, P. N. (1999). Deductive reasoning. *Annual Review of Psychology*, *50*, 109–135.

Kagehiro, D. K., & Stanton, W. C. (1985). Legal vs. quantified definitions of standards of proof. *Law and Human Behavior*, *9*, 159–178.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*, 697–720.

Kahneman, D., & Tversky, A. (1982). The psychology of preferences. *Scientific American*, *246*, 160–173.

Kempf-Leonard, K., & Sample, L. L. (2001). Have federal sentencing guidelines reduced severity? An examination of one circuit. *Journal of Quantitative Criminology*, *17*, 111–144.

Kerr, N. L. (1978). Severity of prescribed penalty and mock juror verdicts. *Journal of Personality and Social Psychology*, *36*, 1431–1442.

Kerr, N. L., Atkins, R. S., Stasser, G., Meek, D., Holt, R. W., & Davis, J. H. (1976). Guilty beyond a reasonable doubt: Effects of concept definition and assigned decision rule on the judgments of mock jurors. *Journal of Personality and Social Psychology*, *6*, 282–294.

Klein, G. (1998). *Sources of power: How people make decisions*. Cambridge, MA: MIT Press.

Kleinmuntz, D. N., & Schkade, D. A. (1993). Information displays and decision processes. *Psychological Science*, *4*, 221–227.

Koch, C. M., & Devine, D. J. (1999). Effects of reasonable doubt definition and inclusion of a lesser charge on jury verdicts. *Law and Human Behavior*, *23*, 653–674.

Koneçni, V. J., & Ebbesen, E. B. (1992). Methodological issues in research on legal decision-making, with special reference to experimental simulations. In F. Lösel, D. Bender, & T. Bliesener (Eds.), *Psychology and law: International perspectives* (pp. 413–423). Berlin: De Gruyter.

Kritzer, H. M. (Ed.). (2002). *Legal systems of the world: A political, social, and cultural encyclopedia* (Vol. 3). Santa Barbara, CA: ABC-CLIO.

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin*, *74*, 100–109.

Lovegrove, A. (1999). Theoretical and methodological issues in the psychological study of judicial sentencing. *Psychology, Crime, and Law*, *5*, 217–250.

MacCoun, R. J., & Kerr, N. L. (1988). Asymmetric influence in mock deliberation: Jurors' bias for leniency. *Journal of Personality and Social Psychology*, *54*, 21–33.

Manktelow, K. (1999). *Reasoning and thinking*. Hove: Psychology Press.

McNicol, D. (1972). *A primer of signal detection theory*. London: Allen and Unwin.

Mello, M., & Robson, R. (1985). Judge over jury: Florida's practice of imposing death over life in capital cases. *Florida State University Law Review*, *13*, 31–75.

Nagel, S. S., Lamm, D., & Neef, M. G. (1981). Decision theory and juror decision-making. In B. D. Sales (Ed.), *Perspectives in law and psychology: The trial process* (Vol. 2, pp. 353–386). Lexington, KY: Lexington Books.

Nijboer, J. F. (1999). Criminal justice system. In J. M. J. Chorus, P. H. M. Gerver, E. H. Hondius, & A. K. Koekkoek (Eds.), *Introduction to Dutch law* (pp. 383–433). Den Haag: Kluwer.

Oakhill, J., & Garnham, A. (1993). On theories of belief bias in syllogictic reasoning. *Cognition*, *46*, 87–92.

Poor, H. V. (1994). *An introduction to signal detection and estimation* (2nd ed.). New York: Springer.

Saks, M. J. (1997). What do jury experiments tell us about how juries (should) make decisions? *Southern California Interdisciplinary Law Journal*, *6*, 1–53.

Sandefur, T. (2003). In defense of plea bargaining: The practice is flawed, but not unconstitutional. *Regulation*, *26*(fall), 28–31.

Seta, J. J., McElroy, T., & Seta, C. E. (2001). To do or not to do: Desirability and consistency mediate judgments of regret. *Journal of Personality and Social Psychology*, *80*, 861–870.

Shafir, E., & LeBoeuf, R. A. (2002). Rationality. *Annual Review of Psychology*, *53*, 491–517.

Shanley, M. G. (1985). *Comparative negligence and jury behavior*. Santa Monica, CA: Rand.

Sheppard, S. (2003). The metamorphoses of reasonable doubt: How changes in the burden of proof have weakened the presumption of innocence. *Notre Dame Law Review*, *78*, 1165–1249.

Simon, R. J. (1969). Judges' translations of burden of proof into statements of probability. *Trial Lawyer's Guide*, *13*, 103–114.

Simon, R. J., & Mahan, L. (1971). Quantifying burdens of proof: A view from the bench, the jury, and the classroom. *Law and Society Review*, *5*, 319–330.

Solan, L. M. (1999). Refocusing the burden of proof in criminal cases: Some doubt about reasonable doubt. *Texas Law Review*, *78*, 105–147.

Stasser, G., Kerr, N. L., & Bray, R. M. (1982). The social psychology of jury deliberations: Structure, process, and product. In N. L. Kerr & R. M. Bray (Eds.), *The psychology of the courtroom* (pp. 221–256). New York: Academic.

Stoffelmayr, E., & Diamond, S. S. (2000). The conflict between precision and flexibility in explaining beyond a reasonable doubt. *Psychology, Public Policy, and Law*, *6*, 769–787.

Sundby, S. E. (1989). The reasonable doubt rule and the meaning of innocence. *Hastings Law Journal*, *40*, 457–510.

Taekema, S. (Ed.). (2004). *Understanding Dutch law*. Den Haag: Boom.

Tak, P. J. (1997). Sentencing and punishment in the Netherlands. In M. Tonry & K. Hatlestad (Eds.), *Sentencing reform in overcrowded times: A comparative perspective* (pp. 194–200). New York: Oxford University Press.

Tak, P. J. P. (2003). *The Dutch criminal justice system: Organisation and operation* (2nd ed.). Den Haag: Wetenschappelijk Onderzoek- en Documentatiecentrum.

Thomas, E., & Parpal, M. (1987). Liability as a function of plaintiff and defendant fault. *Journal of Personality and Social Psychology*, *53*, 843–857.

Tindale, R. S., Davis, J. H., Vollrath, D. A., Nagao, D. H., & Hinsz, V. B. (1990). Asymmetrical social influence in freely interacting groups: A test of three models. *Journal of Personality and Social Psychology*, *58*, 438–449.

Todd, P. M., & Gigerenzer, G. (2000). Précis of simple heuristics that make us smart. *Behavioral and Brain Sciences*, *23*, 727–741.

Tonry, M. & Hatlestad, K. (Eds.). (1997). *Sentencing reform in overcrowded times: A comparative perspective*. New York: Oxford University Press.

Twining, W. L. (1991). *Rethinking evidence*. Evanston, IL: Northwestern University Press.

Twining, W. L. (1995). Anchored narratives: A comment. *European Journal of Crime, Criminal Law and Criminal Justice*, *1995*, 106–114.

Ulmer, J. T., & Kramer, J. H. (1998). The use and transformation of formal decision-making criteria: Sentencing guidelines, organizational contexts, and case processing strategies. *Social Problems*, *45*, 248–267.

van Koppen, P. J. (2002). The Netherlands. In H. M. Kritzer (Ed.), *Legal systems of the world: A political, social, and cultural encyclopedia* (Vol. III, pp. 1114–1121). Santa Barbara, CA: ABC-CLIO.

Vidmar, N. J., Lee, J., Cohen, E., & Stewart, A. (1994). Damage awards and jurors' responsibility ascriptions in medical versus automobile negligence cases. *Behavioral Sciences and the Law*, *12*, 149–160.

Viscusi, W. K. (1988). Pain and suffering in product liability cases: Systematic compensation or capricious awards? *International Review of Law and Economics*, *8*, 203–220.

Volokh, A. (1997). Guilty men. *University of Pennsylvania Law Review*, *146*, 173–211.

Vrij, A. (2003). We will protect your wife and child, but only if you confess: Police interrogations in England and the Netherlands. In P. J. van Koppen & S. D. Penrod (Eds.), *Adversarial versus inquisitorial justice: Psychological perspectives on criminal justice systems* (pp. 53–80). New York: Plenum.

Wagenaar, W. A., van Koppen, P. J., & Crombag, H. F. M. (1993). *Anchored narratives: The psychology of criminal evidence*. London: Harvester Wheatsheaf.

Wald, A. (1947). *Sequental analysis*. New York: Wiley.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (pp. 135–151). Harmondsworth: Penguin.

Wigmore, J. H. (1970). *Evidence* (3rd revised ed. by J. H. Chadbourn). Boston: Little Brown.

Wissler, R. L., Evans, D. L., Hart, A. J., Morry, M. M., & Saks, M. J. (1997). Explaining 'pain and suffering' awards: The role of injury characteristics and fault attributions. *Law and Human Behavior*, *21*, 181–207.

Wissler, R. L., Kuehn, P. F., & Saks, M. J. (2000). Instructing jurors on general damages in personal injury cases: Problems and possibilities. *Psychology, Public Policy, and Law*, *6*, 712–742.

Zeelenberg, M. (1999). Anticipated regret, expected feedback and behavioral decision making. *Journal of Behavioral Decision Making*, *12*, 93–106.

Zeelenberg, M., van den Bos, K., van Dijk, E., & Pieters, R. (2002). The inaction effect in the psychology of regret. *Journal of Personality and Social Psychology*, *82*, 314–327.

Received 13 December 2005; revised version received 22 May 2006