Article

Reasons *Pro et Contra* as a Debiasing Technique in Legal Contexts

Frank Zenker
Department of Philosophy and Cognitive Science,
Lund University, Lund, Sweden; Slovak Academy of
Sciences, Institute of Philosophy, Bratislava, Slovakia;
Department of Philosophy, Konstanz University,
Konstanz, Germany

Christian Dahlman
Law Faculty, Lund University, Lund, Sweden

Rasmus Bååth
Department of Philosophy and Cognitive Science, Lund University, Lund, Sweden

Farhan Sarwar
Department of Psychology, Lund University, Lund, Sweden

Abstract

Although legal contexts are subject to biased reasoning and decision making, to identify and test debiasing techniques has largely remained an open task. We report on experimentally deploying the technique "giving reasons *pro et contra*" with professional ($N = 239$) and lay judges ($N = 372$) at Swedish municipal courts. Using a mock legal scenario, participants assessed the relevance of an eyewitness's previous conviction for his credibility. On average, both groups displayed low degrees of bias. We observed a small positive debiasing effect only for professional judges. Strong evidence was obtained for a relation between profession and relevance-assessment: Lay judges seemed to assign a greater importance to the prior conviction than professional judges did. We discuss challenges for future research, calling other research groups to contribute additional samples.

Keywords

Corresponding Author:
Frank Zenker, Lund University, LUX, Box 192, Lund 221 00, Sweden.
Email: frank.zenker@fil.lu.se

Introduction

Many professional judges assume that (i) non-jurist decision makers regularly err in assessing the relevance of evidence, whereas (ii) judges mostly avoid such error. For some five decades, however, research on heuristics and biases has supported (i) also for judges, thus undermining (ii). Within and between (groups of) agents, therefore, relevance-assessments may differ for intuitive and deliberative modes of reasoning (see e.g., Frenkel & Stark, 2015, esp 8–15; Langevoort, 1998; cf. Mitchell, 2002). Biased decision making is thus (rightly) thought to occur also in legal contexts. That it ought to be reduced requires no argument. Rather, empirical knowledge is wanted how to do this reliably.

Our research addresses four related questions by way of experimentation and interpretative analysis: (1) What is the accuracy-difference between judges' and laypersons' assessments of the relevance of legal evidence (or: Are judges better at activating "system two")? (2) Do relevance-assessments improve in response to deploying a debiasing technique? (3) What is the optimal allocation between debiasing techniques and biases? (4) How to improve debiasing techniques?

Focusing on the first two questions, we report on a pilot-study with Swedish professional judges and lay judges[1] who assessed a written mock legal scenario containing bias-triggering information. Unlike participants in the control group, experimental group-members were instructed "to give reasons pro/con" before stating their assessment. Assessing the effect of this intervention thus contributes to evaluating its potential in (re-)aligning behavior with a normative standard.

We introduce basics on biases and debiasing in the next section, and then the method, its main result, offer a discussion, and finally state our conclusions.

Biases and debiasing

What authors such as Kahneman and Tverksy (1982, 1996) or Kahneman (2011) call *biases*, philosophers and law scholars normally associate with the *fallacies*. After all, both fields share an Aristotelian tradition, specifically its critique of (Sophistic) audience persuasion. Among those carrying this tradition into the modern age are the 16th century Francis Bacon delivering his *idolatry*, the 17th century John Locke, and the 18th century Jeremy Bentham, Richard Whately, and John Stuart Mill (see Hansen, 2015). Since Hamblin (1970), fallacies are standard research objects for speech communication, rhetoric, and argumentation studies, among others. Notably, the interpretation of fallacies as *reasoning errors* was there severed from fallacies as *problematic arguments* (e.g., van Eemeren & Grootendorst, 1984). Most psychologist and cognitive scientists, by contrast, endorse the first interpretation.

Although many empirical studies support the assumed operation of biases in individuals and groups, few studies pertain to the legal context. Exceptions are, among others, Guthrie, Rachlinski, and Wistrich's (2007) study of anchoring, hindsight bias and base rate neglect, and English, Mussweiler, and Strack's (2006) study of anchoring. Both support that biases influence legal decision making (see Zenker & Dahlman, 2016a, for further references).

Biases are generally *latent*—subjects tend to be unaware of them. As extant research suggests, the primary challenge in applying a debiasing technique (especially in self-application) is to suspend latency (Kahneman, 2011; Kenyon, 2014; Pronin & Kugler, 2007; Pronin, Lin, & Ross, 2002; Willingham, 2007). By definition, a technique successfully debiases if it brings forth a decision that qualitatively differs from what deploying a heuristic[2] yields, but also complies with a normative standard (e.g., positive law).

Extant research also identifies a number of debiasing techniques for the legal context (e.g., Guthrie et al., 2007; Irwin & Daniel, 2010). Their underlying principles are sometimes incorporated into, or indeed originate with, procedural or substantial law (see Zenker & Dahlman, 2016b). These techniques included the following:

- *Accountability*: Legal decisions are subject to review by higher courts (Arkes, 1991).
- *Devil's advocate*: Reminding subjects of the hypothetical possibility of the opposite standpoint (Lord, Lepper, & Preston, 1984; Mussweiler, Strack, & Pfeifer, 2000).
- *Giving reasons* (Hodgkinson et al., 1999; Koriat, Lichtenstein, & Fischhoff, 1980; Larrick, 2004, p. 323; Mumma & Wilson, 1995).
- *Censorship*: When evidence counts as inadmissible, this may avoid biases triggered by such evidence.
- *Reducing discretion*: Formulating legal norms that leave less room for a judge's interpretation (e.g., explicit checklists or a pre-set damage amount).

A number of studies suggest that providing incentives and time for reasoning (including its moral variant) can help override intuitive responses (e.g., Paxton et al., 2012). In legal contexts, the potential debiasing effect of the *obligation to give reasons for a judgment* is known as the "it won't write phenomenon." Here, assessment that seemed sound "in the head" may strike the judge as unbalanced when she writes it out (Cohen, 2015; Merrill, 1980; Posner, 1995; Waits, 1983). Studies on the benefits of written versus oral reasoning, however, are inconclusive, leaving the optimal mode for each type of legal case unknown (Oldfather, 2007).

Zenker and Dahlman (2016a) review research on debiasing in legal contexts, including key methodological issues and additional references. They argue that successful debiasing techniques should address aspects of *cognition*, *motivation*,

and *technology*. For a given technique needs to raise awareness of the bias (cognition) in ways that sustain or increase an agent's impetus to avoid biased reasoning (motivation), while providing information she can in fact deploy to correct extant reasoning (technology). Generally, the effects such techniques induce should generate decisions that remain within the law.

The present study focuses on one aspect: cognition. Empirically examining a debiasing technique in view of a bias-triggering mock scenario here assesses the extent to which a hypothetical (yet realistic) legal decision may be subject to biases (if judges' and laypersons' decisions "in the lab" are representative of behavior "outside"). This estimates the potential of explicit instructions to mitigate biases (if what works in the lab indicates that it succeeds outside), and in the long run yields information on the best way for decision makers to deploy a given technique.

Method

To investigate whether giving reasons *pro et contra* has a debiasing effect, we provided two groups of experimental participants with a scenario containing bias-triggering information. A pen-and-paper questionnaire instructed members of the experimental (or debias) group to give reasons *pro et contra* before stating their answers; the control group went ahead without such instruction. Randomly assigned to a group, participants were asked to answer personally rather than delegate (e.g., to a clerk). This design specifically investigates if the instruction to give reasons has a debiasing effect.

Our scenario describes an adult—referred to as "Tony T"—who testifies as a witness in a criminal trial. The focal question regards the extent (if any) to which his being a convicted felon affects his credibility as a witness. Such character evidence may trigger a bias known as "devil effect" or "reverse halo effect" (Thorndike, 1920). Here, a negative personal fact (the prior conviction) is assigned exaggerated importance when judging a personal feature (credibility as witness). The rich literature on this effect includes Davies (1991), Tillers (1997), Cook, Marsh, and Hicks (2003), Hunt and Budsheim (2004), Walton (2006), and Redmayne (2015).

Although character evidence potentially triggers a devil effect, an alternative scenario could of course trigger another bias. Rather than investigate character evidence or the halo/devil effect itself, however, we addressed whether giving reasons *pro et contra* has a debiasing effect. We did not a priori assume that it necessarily instantiates a bias if the prior conviction negatively affects the witness's credibility. Rather, we took a bias to be clearly instantiated if the assessed relevance of the witness's prior conviction for his trustworthiness statistically significantly differs between control and experimental group.

Using a between-subjects design, we sent a personal letter to *all* 667 professional judges at municipal courts in Sweden, asking them to return our

anonymous pen-and-paper questionnaire. By way of the court's chief judge, we similarly asked 738 lay judges to assess what we generally call *prior conviction relevance* (PCR), as operationalized in the following mock scenario:

> Sebastian P is charged for assault. According to the prosecutor's charge, Sebastian P assaulted Victor A, on July 20, 2012 at 23:30 outside a cinema in central Malmö, by repeated blows to the head. Sebastian P testifies that he acted in self-defense and denies the charges. One of the witnesses in the trial is Tony T, who was at the site on that particular evening. During the examination of the witness Tony T, it emerges that he had recently served a two-year prison sentence for illegal possession of weapons and arms trafficking.
>
> Which of the following best describes your assessment? (Tick one option only.)
>
> - Tony T's previous conviction for illegal possession of weapons and arms trafficking affects the assessment of his credibility as a witness in the current trial. When various factors are weighed, the fact that he had previously been convicted of illegal possession of weapons and arms trafficking is strongly to his disadvantage.
>
> - (as above) . . . is clearly to his disadvantage.
>
> - (as above) . . . is somewhat to his disadvantage.
>
> - Tony T's previous conviction for illegal possession of weapons and arms trafficking does not affect the assessment of his credibility as a witness in the current trial.

Professional and lay judges in the experimental group were asked—after presenting the scenario but before the focal question and alternative answers—to state reasons *both* why Tony T's prior conviction would *and* why it would not affect his credibility in the present case. No instructions were given in the control group. We coded responses on a four-point ordinal scale as *not relevant*, *somewhat*, *clearly*, and *strongly to the witness's disadvantage*.

Totally, 239 professional judges (40% response rate) answered the questionnaire, 143 of which (59.8% of sample) did *not* receive debiasing instruction (control group). Another 96 participants (40.2% of sample) were instructed to give pro/con-reasons before stating their assessment (debiasing group); 372 lay judges (52% response rate) also answered the questionnaire, of which 171 (45.9%) belonged to the experimental and 201 (54.1%) to the control group. The response rate is unbalanced since participants were free to return the questionnaire (see Discussion section). We excluded experimental group members who did not state any pro/con-reasons. No other manipulation or exclusion occurred; participants did not receive compensation.

Typical responses from both samples include the following *pro/con-reasons*:

Prior conviction is relevant (pro)

- Tony T lacks a barrier to breaking the law
- Tony T may have an interest (e.g., revenge)
- Tony T commands reduced "citizenship-capital"
- Tony T has a pro-attitude to violence

Prior conviction is not relevant (con)

- Unrelated event/circumstances
- No evidence that prior conviction matters
- Prior conviction *should* be irrelevant
- Current testimony occurs under oath

Conducting exploratory research to estimate parameters, we did not formulate a point-hypothesis to code the normatively correct response prior to deploying the questionnaire. But we expected that participants would judge the prior conviction to have some negative relevance effect on credibility, a judgment that should be less pronounced in the debiasing group. So we did not simply assume the presence of a bias if prior conviction negatively affects the witness's credibility. Rather, we took a bias to be present if control and experimental group participants arrive at significantly different assessments of PCR. Specifically, we assumed that "giving reasons *pro et contra*" induces a debiasing effect, if the experimental group displays a lower average assessment of PCR.

Results

We first describe data from professional judges. Fewer participants in the debiasing than in the control group took the witness's previous conviction to be *clearly* or *strongly* to his disadvantage in the present case, namely six and respectively one (4.2% and 0.7% of sample) versus zero participants. This provides a weak reason to maintain that the technique had an ameliorating effect on judges. Moreover, 28 judges in the control group (19.6% of group) found the witness's prior conviction *somewhat* negatively relevant. Finally, 20 judges in the experimental group (12.8% of group) so register *despite* the technique being deployed.

Turning to lay judges, a noteworthy difference between control and experimental group was not observed: 7% and 8% of lay judges found the prior conviction *clearly* or, respectively, *strongly relevant*; 30% in each group found it *somewhat relevant*; 61% and 63%, respectively, found it *not relevant* (see Table 1 and Figure 1). The overall effect of deploying the technique "giving reasons *pro et contra*" thus was *prima facie* miniscule.

Table 1. Responses from Swedish judges and lay judges (*N* = number of subjects; all percentages rounded; values <2 rounded to first decimal).

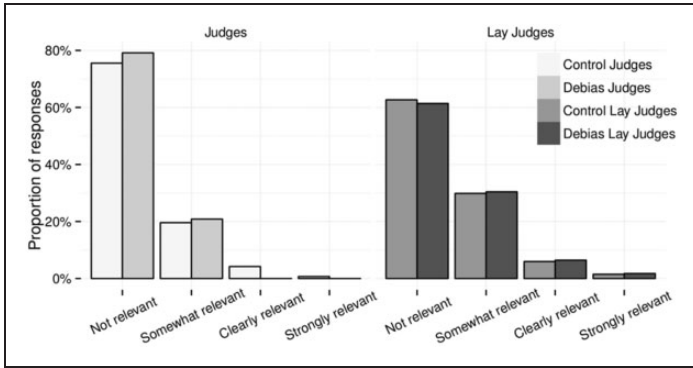|		Not relevant	Somewhat relevant	Clearly relevant	Strongly relevant	N
Judges	Control	108 (76%)	28 (20%)	6 (4%)	1 (.7%)	143
	Debias	76 (79%)	20 (21%)	0 (0%)	0 (0%)	96
	Total	184 (77%)	48 (20%)	6 (3%)	1 (.4%)	239
Lay Judges	Control	126 (63%)	60 (30%)	12 (6%)	3 (1.4%)	201
	Debias	105 (61%)	52 (30%)	11 (6%)	2 (1.1%)	171
	Total	231 (62%)	112 (30%)	23 (6%)	5 (1.3%)	372

Figure 1. Proportion of responses from judges and lay judges with respect to prior conviction relevance in the Tony T scenario.

To quantify differences between the control and experimental groups of professional and lay judges, we subjected data to ordered probit analysis.[3] This assumes that underlying the *ordinal* measurement scale for responses is a *continuous* random variable representing participants' PCR-assessment. Although the value of this latent PCR-variable has no direct interpretation, it nevertheless provides a relative measure of PCR—where a higher value implies that the prior conviction is more relevant. It is crucial for our statistical analysis that the expected PCR-value measures the *group's sentiment*, so as to compare groups.

Using maximum likelihood-estimation, we gauged the parameters of the PCR-variable to yield estimates under which the ordered probit model is most likely to generate data in Table 1. In virtue of being maximally consistent with data, we can interpret this hypothetical model as the *most probable continuous distribution of the latent PCR-variable* among respondents (see Figure 2).
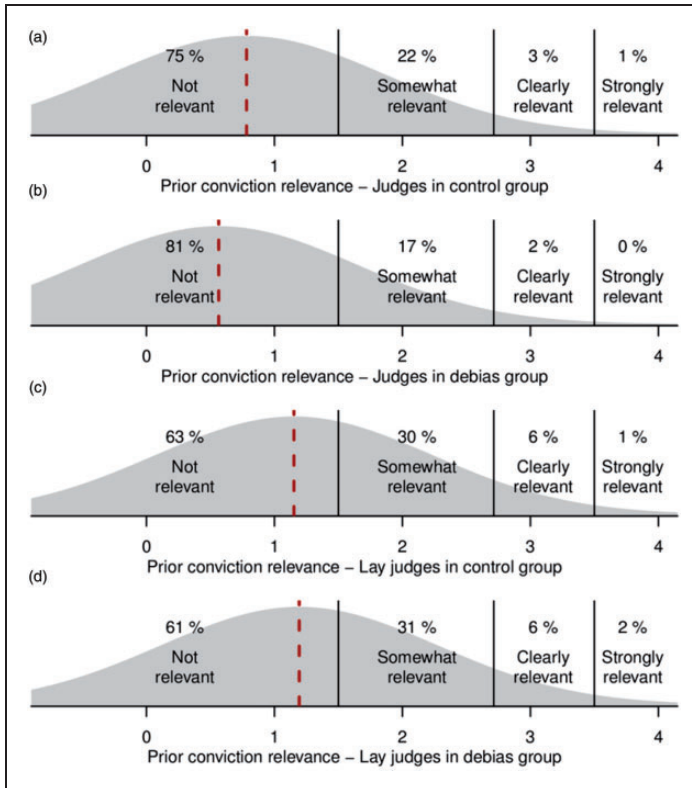
Figure 2. Probability distribution of latent "prior conviction relevance"-variable (PCR) for judges and lay judges in debiasing and control group (percentages rounded to nearest integer).

The shaded curve in Figure 2 represents the maximum likelihood estimate of the latent PCR-variable among judges and lay judges. Each of the four regions in panels A to D corresponds to a possible questionnaire-response. The percentage of the area corresponding to (the part of this curve crossing) a region states the model's probability estimate that group-members (as a collective) give this response. With dashed vertical lines indicating the expected value of the PCR-variable, the displacement of the PCR-value thus marks the debiasing technique's impact.

Comparing panels A to B and C to D of Figure 2, there is a visible difference in PCR assessment between the debiasing and the control group of judges. But hardly any difference is observed for lay judges. However, there is a substantial, and noteworthy, difference insofar as professional judges viewed the prior conviction as *less relevant* than lay judges did.
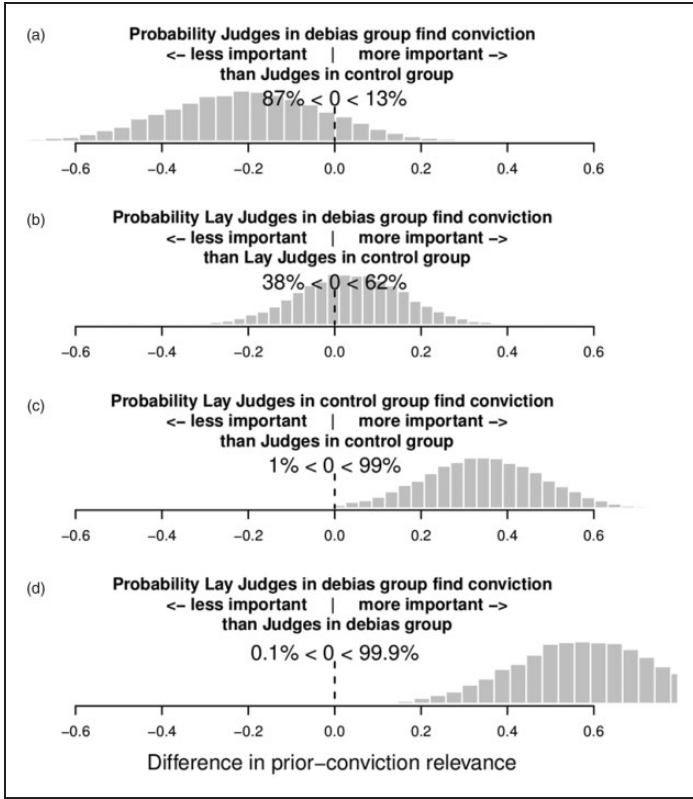
Figure 3. Distribution of probabilities given the ordered probit model and the data from professional and lay judges.

A Bayesian analysis gauged the uncertainty in the estimates obtained from ordered probit analysis, thus quantifying how consistent aggregated data are with the hypothesis that "giving reasons pro/con" had an ameliorating effect (Figure 3).[4] Figure 3 shows the probable differences in the expected PCR-value for all four groups. Given model and data, in the debiasing group, we obtain an 87% probability that *judges*, and a 38% probability that *lay judges* found the prior conviction less relevant than their peers in the control groups (Figure 3, panels A, B).

Comparing judges and lay judges in the control and debiasing group (Figure 3, panels C, D), moreover, there is a 99% probability that lay judges assigned a higher PCR compared to judges, where evidence from the debiasing group registers slightly stronger.

As an alternative technique to ordered probit analysis, we subjected data to a 2×2 analysis of variance test. The first factor was the profession (professional

vs. lay judges), the second the control versus debias-condition. We observed a highly significant main effect of profession, $F(3, 610) = 17.37$, $p < .0001$, partial eta^2 = .03, observed power = .99; lay judges: Mean = .47, SD = .67; professional judges: Mean = .26, SD = .52.

This analysis provides *weak* evidence that deploying the technique had a positive debiasing effect on judges, but not on lay judges, and *strong* evidence that lay judges assigned a higher PCR than professional judges.

Discussion

In this study, professional and lay judges displayed low degrees of bias. Our experimental data did not yield strong evidence for a debiasing effect of the technique "giving reasons pro/con" onto participants' responses. Rather, an 87.1% probability of a bias-ameliorating effect is at best weak evidence. Overall, lay judges assigned *greater* weight than professional judges to the witness's previous conviction. Moreover, and perhaps disturbingly, lay judges in the debiasing group displayed an *increased* mean score compared to the respective control group. This does not amount to a causal interpretation, of course. But differences between professional judges' and lay judges' training and work-experience plausibly account for this interaction effect.

Momentarily restricting discussion to data from professional judges, around 60% of control group participants returned the questionnaire, while some 40% of experimental group participants did (see Method section). This imbalanced response rate potentially lets data bear an attenuation effect. Speculatively, since answering the focal question takes time, the more a judge is pressed for it, the less likely she would be to return the questionnaire. On the additional assumption that a senior judge is more severely pressed for time than a junior colleague, data might therefore relatively over-represent junior judges' responses. A related assumption is that the intervention was more effective among senior than junior judges. So results might indicate that junior colleagues are comparatively less likely to successfully debias. Finally, if a more cautious decision maker were more likely *not* to return the questionnaire than a less cautions one, a similar heterogeneity issue arises. (Any inference from a heterogeneous sample, of course, must be qualified accordingly.)

After the fact, however, there is no telling. Our anonymous questionnaire keeps us from reporting relevant information. Future work should control individual and demographic differences between respondents that bear on data interpretation. Pace the caveats, the Tony T mock case did *not* induce a strong bias among professional or lay judges. By and large, professional judges assigned merely *some* weight to the previous conviction, while lay judges assigned a greater weight.

Although "giving reasons *pro et contra*" did not meet with a strongly biased sample, the technique does appear to "take off the edge." After all, the number

of extreme judgments among professional judges in the debiasing group is reduced vis-à-vis the control group. Removing but one extreme judgment may already be an important and desirable outcome, but it remains a small effect. (Whether this holds equally for each group member is again subject to the above caveats.)

Unexpectedly, when deploying the technique among the comparatively more biased sample of lay judges, it not only failed to mitigate, but comparatively slightly "worsened" the group's overall judgment. Since the statistical evidence was very weak, however, we cannot easily ascribe this effect directly to the technique, $F(3, 610) = .44$, $p < .51$, partial eta$^2 = .001$, observed power $= .10$.

A relevant concern is that it takes additional data to achieve greater certainty as to whether a debiasing effect arises under our experimental set-up. But consider that the sample of professional judges ($n = 239$) already comprises 40% of the relevant national *population*. For formal reasons alone, of course, before a small effect can register as statistically significant, one must collected a sufficiently large sample. But to increase *this* sample presents obvious difficulties.

In terms of substance, application and training, moreover, legal systems have genuinely national characteristics. So completing the sample with data from judges at courts *other than* Swedish ones might seem to incur special challenges. But we grant that differences in national law are negligible regarding the question whether a previous conviction negatively affects an eyewitness's trustworthiness *generally*.

Technical difficulties, by contrast, do not arise, since individually underpowered studies can be meaningfully aggregated. So one can "make up" for a small sample (see Witte & Zenker, 2016a, 2016b; cf. Marsman, Ly, & Wagenmakers, 2016). Future research, therefore, can contribute additional samples.

At the same time, our discussion reminds of a conundrum: There may be biases whose presence, and debiasing techniques whose effect, one cannot demonstrate by obtaining strong experimental evidence for a significant difference between control and debiasing groups, namely when the effect is too small to yield substantial evidence even in the *population* of Swedish judges.

To explain this, we rely on a 2×2 analysis of variance test. The effect on responses in debiasing and control groups was statistically non-significant for the Tony T case, $F(3, 610) = .44$, $p < .51$, partial eta$^2 = .001$, observed power $= .10$. It follows that, other things being equal, registering this small an effect as a statistically significant deviation from random (power $= .95$; alpha-error $= .05$) requires a staggering $N = 12,994,712$. For small populations, hence, a challenge remains that experimental conditions must trigger *stronger* biases.

It may therefore strike readers less unfamiliar with experimental work as a negative that we cannot say if "giving reasons *pro et contra*" has a debiasing effect in legal contexts. But such is the nature of explorative research. We might add that so-called "inconclusive" results have their rightful place. In fact,

hiding similar results in the file-drawer would risk biasing meta-analyses by positive results.

Conclusion

In our sample of judges and lay judges at Swedish municipal courts, the Tony T mock legal scenario failed to meet "sufficiently biased" respondents. Rather few experimental participants assigned any greater relevance to the witness's prior conviction for his credibility in the present case. Thus, the debiasing technique "giving reasons *pro et contra*" merely produced a rather small positive effect.

Although our main result is therefore inconclusive, it provides weak evidence for the technique's effectiveness among professional judges. Results differed in the normatively opposite direction among lay judges, however, who were slightly more biased than professional judges. Moreover, the technique may have had a *slightly adverse* effect: Lay judges assigned a somewhat increased weight to the relevance of the witness' previous conviction. But this interpretation is subject to caveats because the effect's direction is uncertain.

Among all measures, we obtained very strong evidence only for the presence of a relation between profession and level of biasedness. The probability was greater than 99% that lay judges are relevantly more biased than professional judges. In generating more substantial evidence for the effectiveness of a debiasing technique, future research should trigger strong(er) biases. We encourage others to adopt our set-up using samples other than judges at Swedish courts.

Notes

1. In the Swedish legal system, criminal cases are decided by a *mixed tribunal* composed of professional judges and lay judges (*nämndemän*); there is no all-citizen jury as in the English tradition. Professional judges are trained in law (LLM degree from a Swedish University), holding permanent positions as magistrates. Lay judges, by contrast, lack legal education. Regional parliaments elect them to sit on a number of trials for a four-year term. Tribunals at municipal courts are is usually composed of a professional judge, who acts as chair, and four lay judges. At appeals courts, by contrast, professional judges are in the majority; a typical tribunal consists of three professional and two lay judges. Swedish procedural code assigns one vote to each professional or lay judge. In practice, professional judges enjoy considerable authority; lay judges tend to follow their judgment (see Hans, 2008, esp. 289).

2. An outcome O_1 of a heuristic reasoning mode H need not differ from an outcome O_2 brought about by reasoning *not* grounded in H. Indeed, O_1 and O_2 may be the same (see e.g., Gigerenzer & Brighton, 2009). Although these outcomes may be unproblematically observable, the processes (heuristic or other) generating them are not. So the "technique T successfully debiases" is an *empirical* statement, only if the outcome T induces differs in substance from the outcome H induces.

3. Analysis relied on the R-statistical environment, using the polr function of the MASS package (Venables & Ripley, 2002). See Daykin and Moffat (2002) for the advantages of paradigmatic applications of ordered probit analysis over linear regression analyses. For instance, ordered probit analysis is not open to the objection that distances between ordinal data points are implicitly treated as being equal.

4. Analysis relied on the R-statistical environment using the MCMCoprobit function in the MCMC package (Martin et al., 2011). We used default priors of the MCMCoprobit function, i.e., non-informative uniform priors over all parameters.

References

Arkes, H. R. (1991). Costs and benefits of judgement errors: Implications for debiasing. *Psychological Bulletin*, *110*, 486–498.

Cohen, M. (2015). When judges have reasons not to give reasons: A comparative law approach. *Washington and Lee Law Review*, *72*, 483–571.

Cook, G., Marsh, R., & Hicks, J. (2003). Halo and devil effects demonstrate valence-based influences on source-monitoring decisions. *Consciousness and Cognition*, *12*, 257–278.

Davies, S. M. (1991). Evidence of character to prove conduct. *Criminal Law Bulletin*, *27*, 504–537.

Daykin, A. R., & Moffat, P. G. (2002). Analyzing ordered responses: A review of the ordered probit model. *Understanding Statistics*, *1*(3), 157–166.

English, B., Mussweiler, T., & Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin*, *32*(2), 188–200.

Frenkel, D. N., & Stark, J. H. (2015). Improving lawyers' judgment: Is mediation training de-biasing. *Harvard Negotiation Law Review*, *21*, 1–58.

Gigerenzer, G., & Brighton, H. (2009). Why biased minds make better inferences. *Topics in Cognitive Science*, *1*, 107–143.

Guthrie, C., Rachlinski, J. J., & Wistrich, A. J. (2007). Blinking on the bench: How judges decide cases. *Cornell Law Review*, *1*, 1–44.

Hamblin, C. (1970). *Fallacies*. London, The England: Methuen.

Hans, V. P. (2008). Jury systems around the world. *Annual Review of Law and Social Science*, *4*, 275–297.

Hansen, H. (2015). Fallacies. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Retrieved from <https://plato.stanford.edu/archives/sum2015/entries/fallacies/>

Hodgkinson, G. P., Bown, N. J., Maule, A. J., Glaister, K. W., & Pearman, A. D., (1999). Breaking the frame: An analysis of strategic cognition and decision making under uncertainty. *Strategic Management Journal*, *20*, 977–985.

Hunt, J., & Budesheim, T. (2004). How jurors use and misuse character evidence. *Journal of Applied Psychology*, *89*, 347–361.

Irwin, J., & Daniel, L. R. (2010). Unconscious influences on judicial decision-making. *McGeorge Law Review*, *43*, 1–20.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Strauss and Giroux.

Kahneman, D., & Tversky, A. (1982). On the study of cognitive illusions. *Cognition*, *11*, 1123–1141.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions: A reply to Gigerenzer's critique. *Psychological Review*, *103*, 582–591.

Kenyon, T. (2014). False polarization: Debiasing as applied social epistemology. *Synthese*, *191*(11), 2529–2547.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 107–118.

Langevoort, D. C. (1998). Behavioral theories of judgment and decision making in legal scholarship: A literature review. *Vanderbilt Law Review*, *51*, 1499–1540.

Larrick, R. P. (2004). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making* (pp. 316–337). Oxford: UK.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Judgment*, *47*(6), 1231–1243.

Marsman, M., Ly, A., & Wagenmakers, E.-J. (2016). Four requirements for an acceptable research program. *Basic and Applied Social Psychology*, *38*, 308–312.

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software*, *42*(9), 1–21.

Merrill, C. (1980). Query: Could judges deliver more justice if they wrote more opinions? *Judicature*, *64*, 435.

Mitchell, G. (2002). Why law and economics' perfect rationality should not be traded for behavioral law and economics' equal incompetence. *Georgetown Law Journal*, *91*, 67–167.

Mumma, G. H., & Wilson, S. B. (1995). Procedural debiasing of primacy/anchoring effects in clinical-like judgments. *Journal of Clinical Psychology*, *51*(6), 841–853.

Mussweiler, T., Strack, F., & Pfeifer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin*, *26*(9), 1142–1150.

Oldfather, C. M. (2007). Writing, Cognition, and the Nature of the Judicial Functions. *Georgetown Law Journal*, *96*, 1283–1345.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science*, *36*(1), 163–177.

Posner, R. (1995). Judges' writing styles (and do they matter?). *University of Chicago Law Review*, *62*, 1421.

Pronin, E., & Kugler, M. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology*, *434*, 565–578.

Pronin, E., Lin, D., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self *versus* others. *Personality and Social Psychology Bulletin*, *28*, 369–381.

Redmayne, M. (2015). *Character in the criminal trial*. Oxford, England: Oxford University Press.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*(1), 25–29.

Tillers, P. (1997). What is wrong with character evidence? *Hastings Law Journal*, *49*, 781–834.

van Eemeren, F. H., & Grootendorst, R. (1984). *Speech acts in argumentative discussions: A theoretical model for the analysis of discussions directed towards solving conflicts of opinion*. Amsterdam, The Netherlands: Walter de Gruyter.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York, NY: Springer.

Waits, K. (1983). Values, intuitions, and opinion writing: The judicial process and state court jurisdiction. *University of Illinois Law Review*, 917–976.

Walton, D. (2006). *Character evidence*. Dordrecht, The Netherlands: Springer.

Willingham, D. T. (2007). Critical thinking: Why is it so hard to teach? *American Educator*, *31*(2), 8–19. (Reprinted as: Willingham, D. T. (2008). Critical thinking: Why is it so hard to teach? *Arts Education Policy Review, 109*(4), 21–32.

Witte, E. H., & Zenker, F. (2016a). Reconstructing recent work on macro-social stress as a research program. *Basic and Applied Social Psychology*, *38*(6), 301–307.

Witte, E. H., & Zenker, F. (2016b). Beyond schools—Reply to Marsman, Ly & Wagenmakers. *Basic and Applied Social Psychology*, *38*(6), 313–317.

Zenker, F., & Dahlman, C. (2016a). Reliable debiasing techniques in legal contexts? Weak signals from a darker corner of the social science universe. In F. Paglieri, L. Bonelli & S. Felletti (Eds.), *The psychology of argument: Cognitive approaches to argumentation and persuasion* (pp. 173–196). London, England: College Publications.

Zenker, F., & Dahlman, C. (2016b). Debiasing and rule of law. In E. Feteris, H. Kloosterhuis, J. Plug & C. Smith (Eds.), *Proceedings of the International Conference "Rule of Law"* (pp. 217–229). The Hague, The Netherlands: Eleven International.

Author Biographies

Frank Zenker is a researcher at Lund University in the Department of Philosophy and Cognitive Science and a member of the LEVIC research

group. His main research interests are in the philosophy of science, cognitive science, and social epistemology.

Christian Dahlman is a professor of jurisprudence at Lund University and a director of the cross-disciplinary research group "Law, Evidence and Cognition" (LEVIC). His research interests include legal evidence, Bayesian modeling, and cognitive bias in legal decision-making.

Rasmus Bååth has a PhD in Cognitive Science from Lund University, Sweden. His areas of interest include human and animal perception, with a special focus on time perception and rhythm perception. He has also worked on new Bayesian methods for the analysis of psychological experiments.

Farhan Sarwar holds a PhD in psychology. Currently he is a researcher at the Department of Psychology, Lund University, Sweden. His research interests include eyewitness memory and meta-memory processes, ear witnesses, investigative interviews, lie-detection, false confessions, semantic spaces, judgment and decision making, argumentation, biases, and de-biasing.