Justice Is Less Blind, and Less Legalistic, than We Thought

Author(s): Holger Spamann and Lars Klöhn

Source: *The Journal of Legal Studies*, June 2016, Vol. 45, No. 2 (June 2016), pp. 255-280

Published by: The University of Chicago Press for The University of Chicago Law School

Stable URL: https://www.jstor.org/stable/10.2307/26458532

Justice Is Less Blind, and Less Legalistic, than We Thought: Evidence from an Experiment with Real Judges

Holger Spamann and Lars Klöhn

ABSTRACT

We experimentally investigate the determinants of judicial decisions in a setting resembling real-world judicial decision making. We gave US federal judges 55 minutes to adjudicate a real appeals case from an international tribunal, with minor modifications to accommodate the experimental treatments. The fictitious briefs focused on one easily understandable issue of law. Our 2 × 2 between-subject factorial design crossed a weak precedent and legally irrelevant defendant characteristics. In a survey, law professors predicted that the precedent would have a stronger effect than the defendant characteristics. In actuality, the precedent had no detectable effect on the judges' decisions, whereas the two defendants' affirmance rates differed by 45 percent. Judges' written reasons, on the other hand, did not mention defendant characteristics, focusing instead on the precedent and other legalistic and policy considerations.

1. INTRODUCTION

Judges are supposed to decide cases neutrally according to the law. In their written opinions, judges justify their decisions with reference to authoritative legal sources—precedents and statutes—even when these sources are not clear. In such cases, lawyers have long understood that

HOLGER SPAMANN is Professor of Law, Harvard Law School. LARS KLÖHN is Professor of Law, Humboldt-University Berlin. We thank Ludwig Maximilian University of Munich for financial support; Roland Ramthun of Docustorm for programming the experimental interface; Alex Whiting for suggesting the aiding-and-abetting controversy; Oren Bar-Gill and Jake Gersen for feedback on the survey design; Nancy Gertner, John Manning, Denise Neary, and the Federal Judicial Center for arranging the session at which this experiment was conducted; Hannah Clark, Priya Gupta, and Caleb Wolanek for logistical help; the law professors who answered the survey; and above all the judges for their participation. For comments, we are grateful to Adam Chilton, Jim Greiner, Dan Kahan, Louis Kaplow, Brian Leiter, Janice Nadler, Jeff Rachlinski, Dan Simon, and seminar participants at Harvard Law School, the University of Bonn, the University of Chicago, the

[*Journal of Legal Studies*, vol. 45 (June 2016)]

the full explanation of the decision must include other factors, whether or not the judge states or even perceives them explicitly. As Llewellyn (1940) wrote already 75 years ago, the real question is which other factors come into play and how much they influence the decision (see also, for example, Kennedy 1998; Simon 1998; Leiter 2005; Epstein, Landes, and Posner 2013). Some factors are almost universally accepted as relevant, particularly moral or policy considerations (which are indeed considered legal factors by some writers). By contrast, others are officially shunned. For example, the biography and character of a criminal defendant should not matter for the determination of guilt in the technical legal sense, particularly the definition of the crime, however relevant such factors might be for lay attributions of blame (compare Nadler 2012; Nadler and McDonnell 2012). We have almost no convincing evidence, however, to what extent the legal system achieves this goal.

In this paper, we report findings from an experiment testing the effect of legally relevant and irrelevant factors side by side. Our 2×2 factorial between-subject design preserves key attributes of real-world judicial decision making. Real judges reviewed a full set of legal briefs and materials for almost 1 hour before rendering a decision with written reasons. The case was based on a real appeals case from the International Criminal Court for the Former Yugoslavia (ICTY). The judges were randomly assigned one of two defendants and one of two precedents. One of the precedents weakly favored the defendant's position (obiter dictum), whereas the other weakly disfavored it (on distinguishable facts). The precedent was the focus of the briefs and the judges' written reasons but had no detectable effect on the judges' decisions. By contrast, the second, legally irrelevant variation had a strong effect. Of the judges who reviewed the case of a nationalist, hateful Serb defendant, 87 percent upheld the conviction, as opposed to only 41 percent of the judges who reviewed the case of a conciliatory, regretful Croat defendant. The judges' written reasons show no awareness of this effect. In a survey, law professors had predicted the opposite result, that is, that even the weak precedent should matter more for the outcome than the defendant.

The key features of our design are its realism, particularly the inten-

University of Michigan, the University of Pennsylvania, Yale Law School, and the 2015 conference of the International Society for New Institutional Economics. This research was approved by Harvard University's Committee on the Use of Human Subjects as protocol IRB15-0206 (main experiment) or exempt as protocol IRB15-2844/3074 (professorial surveys).

sive participation of real judges, and the side-by-side comparison of legally relevant and irrelevant factors. The latter feature is crucial because, as noted above, the existence of some extralegal influences on judicial decisions is not in serious dispute. The interesting question is their strength relative to that of the legally relevant factors. In particular, it would hardly be reassuring if one found that a particular irrelevant factor does not matter but neither do the factors that should, in particular those that the judges mention in their written opinions. This is not to say that our experiment is able to test the strength of legally relevant and irrelevant factors in general. We test particular legal factors against particular irrelevant factors. In particular, we test a weak precedent against certain defendant characteristics. We might obtain completely different results if we tested a strong precedent or if we tested it against the racial or political influences that stir most controversy in the United States. We put at least one point on the map, however, and this point fits better with some views of the general landscape than with others. We document explicitly that the result is not consistent with law professors' prior views: our results differ strongly from law professors' predictions, which we elicited in a survey.

Realism matters because specificities of judges and the legal process might well have deep effects on legal reasoning as practiced in courts. General psychology furnishes plenty of reasons to be very skeptical of neutral legal reasoning.[1] Indeed, some realist lawyers such as Frank (1930 [2009]) prominently endorse such skepticism. As other realist lawyers like Llewellyn (1940) point out, however, judges are not merely human—they are humans with a particular ability, training, and expertise, and they are embedded in institutions that might well allow them to do considerably better at objective reasoning tasks than average humans. First, judges are professionals who are highly trained and selected specifically to interpret the legally relevant factors and ignore the irrelevant ones (Spellman and Schauer 2012; Kahan 2015). Second, characteristic features of the judicial process may inoculate judges against the influence of legally

1. We do not purport to differentiate between these reasons; that is, we remain agnostic about any particular psychological mechanisms. To give just one example, if competing objective standards lead to different outcomes, people may unconsciously choose whatever standard justifies their preferred outcomes, which creates merely an illusion of objectivity (Norton, Vandello, and Darley 2004). Similarly, decision makers tend to perceive information in a biased manner once they have developed an initial inclination toward one option (DeKay 2015; Russo 2015), as a judge might after initial exposure to the facts. On motivated reasoning in law generally, see Sood (2013).

irrelevant factors. In particular, accountability induced by the need to give written reasons may reduce the impact of irrelevant factors (Lerner, Goldberg, and Tetlock 1998), albeit not always (Norton, Vandello, and Darley 2004). Similarly, having time to reflect on the issues presented might eliminate hunches driven by legally irrelevant factors (analogous to Paxton, Ungar, and Greene 2012), notwithstanding counterexamples (Schwitzgebel and Cushman 2015). Our results suggest that judges in these conditions ultimately behave much like regular lab participants. It is essential, however, to establish such similarity empirically (Spellman and Schauer 2012; Kahan 2015).

Many experiments have examined heuristics and biases in legal fact finding, damage awards, and sentencing with both judge and lay subjects (for example, Englich, Mussweiler, and Strack 2006; Guthrie, Rachlinski, and Wistrich 2007; Rachlinski et al. 2009; Rachlinski, Wistrich, and Guthrie 2013; Kahan 2010; Kahan et al. 2012; Nadler and McDonnell 2012; Sood and Darley 2012).[2] By contrast, few experiments examine legal reasoning such as statutory interpretation (Holyoak and Simon 1999; Simon et al. 2001; Braman and Nelson 2007; Furgeson, Babcock, and Shane 2008a, 2008b; Wistrich, Rachlinski, and Guthrie 2014; Kahan et al. 2016). The only ones to do so with real judges are Wistrich, Rachlinski, and Guthrie (2014) and Kahan et al. (2016). Unlike our study, however, these two studies provide context for the decision only as a vignette, which creates much less immersion in the case;[3] do not ask for written reasons, which imposes less accountability and makes it impossible to study the judges' subjective reasons; and do not cross vary the law, which makes it impossible to compare the strength of the emotional or political effect with the strength of a legal effect. Kahan (2010) presents the only experiment comparing the effects of legal and extralegal factors (using lay subjects).

With one possible exception, the results of the prior studies are consistent with ours. In particular, Wistrich, Rachlinski, and Guthrie (2014) also find that judges are affected by litigants' valence in cases ostensibly turning on the interpretation of a statute or similar legal questions. Like us, Wistrich, Rachlinski, and Guthrie (2014) randomly assign judges to

2. Kahan (2010), Kahan et al. (2012), Nadler and McDonnell (2012), and Sood and Darley (2012) concern the assessment of legally charged complex facts such as causation or consent, which straddle the boundary between fact and law's application to fact.

3. Presumably, the experiments in Wistrich, Rachlinski, and Guthrie (2014) and Kahan et al. (2016) were also much shorter.

sympathetic or unsympathetic defendants in otherwise identical cases. By contrast, in Kahan et al. (2016), defendants were not sympathetic or unsympathetic per se. Rather, the appeal of the randomly assigned defendants' actions in Kahan et al. (2016)—pro- or anti-immigrant, pro-life or pro-choice—depended on subjects' political predisposition. Kahan et al. (2016) find that lay subjects' and law students' responses varied in accordance with these predispositions. For example, liberal lay subjects were more likely to find a violation by anti-immigrant activists than by pro-immigrant activists. However, judges and lawyers did not exhibit such differential responses. A possible reconciliation of these results with ours and with Wistrich, Rachlinski, and Guthrie (2014) is that judges and lawyers have been trained to disregard cultural-political but not personal implications. Another possibility that would reconcile the results of Kahan et al. (2016) with ours—but not with Wistrich, Rachlinski, and Guthrie (2014)—is that the greater cognitive load in our experiment (and, arguably, in reality) interferes with controlled cognition (compare Greene et al. 2008). The discrepancy remains an important challenge for future research. Finally, we note that differences between judges (as observed on panels or through random assignment to individual cases) demonstrate the importance of factors beyond statutes and precedents (because statutes and precedents are uniform for all judges), but they do not reveal which other factors matter (for example, judicial philosophy versus racial bias) or how their influence compares with that of statutes and precedents.

The paper is structured as follows. Section 2 describes the experiment in detail. Section 3 reports law professors' predictions for the experiment, which we elicited in an online survey. Section 4 presents the results of the experiment. Section 5 discusses implications and ecological validity. Section 6 concludes. The main materials used in the experiment and the associated survey of professors are reproduced in the online appendix.

2. EXPERIMENTAL DESIGN

2.1. General Approach

We briefly outline some high-level design choices before discussing details of the design in the subsequent sections. We defer a detailed discussion of ecological validity to Section 5.2.

Our experiment aimed to study the effects of legally relevant and ir-

relevant factors under conditions that unite the key features of judicial decision making in the real world. For this purpose, we had real judges decide a real case with briefs and legal materials in 1 hour in the setting described in Section 2.2. These conditions correspond roughly to a single judge ruling on a motion without a hearing under severe time constraints. We do not purport to recreate the conditions of high-stakes, long-duration, multijudge proceedings such as those at the US Supreme Court.

To study the effect of legal materials, we needed to be able to vary these materials without arousing suspicion by knowledgeable judges. For this reason, we chose the international case described in Section 2.3 on the assumption—borne out by an exit questionnaire—that the US judges would be unfamiliar with the applicable law. At the same time, the legal question was simple and familiar enough for the judges to understand with ease. We chose an appeals case because appeals cases are limited to legal questions, which are the focus of our study.

The legally relevant factor we study is a weak precedent, as described in Section 2.4.1. This is a suggestive but not binding decision by another court or panel, which tends to form the basis of appellate opinions (or more generally of opinions in cases in which the law is not clear). We did not test the common(-sense) conjecture that a strong precedent (or statute, for that matter) would reduce or even eliminate interpretative leeway and hence the effect of legally irrelevant factors. As regards the legally irrelevant factors, our design cumulated three (nationality, political views, remorse) because we did not know which, if any, might matter and because their much less prominent position in the written materials might lead them to be overlooked in the experiment (as opposed to in a real-world courtroom). We did not attempt to disentangle which of the three legally irrelevant factors had an effect, if any, or through which psychological mechanisms they would do so (compare Spellman 2010). For this reason, our design did not need to address the possibility that small, incidental differences such as the sound of the defendant's name or the layout of the briefs had an effect: from our perspective, all that matters is that these differences are legally irrelevant.

2.2. Setting

We conducted the experiment at a 3-day workshop for US federal judges organized jointly by Harvard Law School and the Federal Judicial Center in April 2015. All participants were US federal judges, including circuit

judges, district judges, bankruptcy judges, and magistrates.[4] The experiment was part of a session called "Behavioral Research on Judicial Decision Making" in the middle of the second morning. Several weeks earlier, the judges had received an invitation to the experiment with all consent-relevant information (online appendix Section A1.1) and a reading assignment: Guthrie, Rachlinski, and Wistrich (2007), which discusses biases in judicial fact finding. The experiment was administered on iPads we provided to the judges.

One of the experimenters welcomed the judges, reminded them of the experiment as described in the invitation letter, and pointed them to the iPads (online appendix Section A1.2). Three student assistants distributed and collected the iPads and were available for help with technical questions; they did not know what the experiment was about.[5] The experimenter stayed in the room but did not interact with the subjects. Participation was voluntary, but all the judges in the room participated in the experiment.[6]

The opening screen on the iPad reminded the judges of the invitation letter, which they could read by clicking on a link. After confirming that they had read the letter and agreed to participate, the judges were shown an instruction page that described their task (online appendix Section A1.3). The instructions invited them to imagine themselves as a judge on the ICTY's appeals chamber reviewing a defendant's appeal of his conviction by the ICTY's trial chamber.[7] The judges were told that they had 50 minutes to reach a decision and submit a brief summary of their reasoning.

When the judges clicked on a button to continue, they were taken to an overview page listing all of the documents available to them (including the instructions), and a clock on the screen started counting down 50 minutes. Besides the instructions, the available documents were a statement of agreed facts (online appendix Section A1.4), briefs for the defen-

4. At the time of the experiment, no circuit judge may have been present. We refrained from collecting this information out of concern for preserving anonymity.

5. The student assistants received few requests for help. The judges seemed comfortable with the technology—almost all arrived at the seminar room with their own tablet or laptop computer.

6. We lost a small number of observations because of technical problems; see the explanation in note 10.

7. The instructions asked the participants "to judge whether the defendant is or is not guilty." This language is technically imprecise because the direct determination of guilt is generally considered a question for the trial court, whereas appeals courts decide whether to reverse or affirm. As we discuss in Section 5.2, we believe this distinction to be without consequence for the experiment and its ecological validity.

dant (appellant) and the prosecution (appellee) (online appendix Sections A1.5 and A1.6, respectively), the ICTY statute, the judgment from the ICTY's trial chamber (roughly 165,000 words), and one precedent from the ICTY's appeals chamber (roughly 37,000 words) that was handed down after the trial judgment in our case. The briefs linked to the most relevant passages in the statute and the precedent. All materials were accessible from a menu on the left of the screen. The long documents had hyperlinked tables of contents.

The briefs and statement of facts each ran under 1,000 words (2 pages). We created these documents from scratch. The instructions recommended reading the briefs and statement of facts in full and consulting the other documents (trial judgment, precedent, statute) as necessary. These other documents were obviously much too long to be read in their entirety in 50 minutes. This was intentional, as real-world judges do not have the time to read all the documents in a case either. However, the most relevant passages of these long documents were referenced and linked from the briefs and could easily be read in this time. Importantly, the legal question in the case was ultimately simple and fully discussed in the short briefs, such that the task was manageable.

A clock on the judges' screens counted down the 50 minutes available, but the judges could choose when to move on to registering their judgment. Some went slightly over time; many finished early. On average, the judges spent 35 minutes (SD = 10 minutes) with the materials before proceeding to judgment.[8] When the judges clicked on the Proceed to Judgment button and confirmed this choice in a pop-up, they were taken to a page that asked them for a tick-the-box answer of guilty or not guilty and, in a text field below, to provide reasons for their decision. Alternatively, the judges could write their reasons on a piece of paper and link it to the rest of their session by noting a randomly generated code on the paper. After the judges submitted and confirmed their judgments, a brief exit questionnaire appeared on the screen. After 55 minutes, the screen prompted the judges to conclude. Several minutes later, the last ones did.

2.3. Legal Context

We derived our case from a real ICTY case, *Prosecutor v. Momčilo Perišić* (case no. IT-04-81). The main question in *Perišić*—in our setup, the only question—was whether a conviction for aiding and abetting under article

8. The mean and standard deviations were calculated using only the observations without the technical issue described in note 10.

7(1) of the ICTY statute requires that the aid be "specifically directed" at the war crime or whether any substantial contribution is sufficient. Defendant Momčilo Perišić had been the highest-ranking general of Yugoslavia for much of the Bosnian war. In this capacity, he had been responsible for organizing various types of Yugoslavian support for the Army of the Republika Srpska (VRS). The VRS was the main armed group of ethnic Serbs in the Bosnian war and committed various war crimes in Bosnia, including the notorious Srebrenica massacre. Yugoslavian support for the VRS included personnel and arms. In 2011, the trial chamber convicted Perišić as an aider and abettor to the VRS crimes (*Prosecutor v. Perišić,* case no. IT-04-81-T, Judgment of the Trial Chamber, September 6, 2011). In a controversial decision from 2013, the ICTY appeals chamber reversed, holding that aiding and abetting required the aid to be "specifically directed" at the crimes (*Prosecutor v. Perišić,* case no. IT-04-81-A, Judgment of the Appeals Chamber, February 28, 2013). Perišić had had knowledge of the VRS war crimes when providing substantial support to the VRS. But the ICTY found that his support was directed merely toward the general war effort of the VRS, not specifically toward its war crimes.

We provided the original *Perišić* trial judgment of the ICTY trial chamber in the materials, except that we changed the date to January 2014 (to make it a live issue), changed the names and some biographical information as described below, and omitted the parts relating to Zagreb. We omitted Zagreb because it proved too difficult to find a credible mirror city targeted by ethnic Croats. We also provided the original ICTY statute and one redacted original precedent from the ICTY appeals chamber, as described below.

We wrote the statement of facts and the briefs with the goal of focusing the judges on only one legal issue, namely, the reach of aiding and abetting liability under article 7(1), as explained above. Toward this goal, we used the title "Statement of Agreed Facts" and began with this sentence: "The parties have agreed that the following key facts are not in dispute." Similarly, the brief for the appellant began with these words: "This appeal concerns a single point of law: whether or not aiding and abetting under Article 7(1) of the Statute governing this Tribunal requires that the assistance be specifically directed to the commission of a crime." Both briefs focused on this issue alone. They discussed the precedent and the policy issues and cited specific passages of the precedent that could be accessed directly using hyperlinks.

2.4. Treatments

We randomly assigned judges to one of the four groups formed by crossing two precedents with two defendants. The randomization mechanism was designed to create groups of equal size.[9] The briefs and the statement of facts were adjusted accordingly.

2.4.1. Precedents. For precedents we used two decisions of the ICTY appeals chamber. In *Prosecutor v. Mitar Vasiljević,* the ICTY appeals chamber had defined aiding and abetting as "specifically directed to assist, encourage or lend moral support to the perpetration of a certain specific crime . . . , and this support has a substantial effect upon the perpetration of the crime" (case no. IT-98-32-A, para. 102, February 25, 2004). This was favorable to our defendant because even the prosecution agreed that he had not specifically directed his support at crimes. By contrast, in *Prosecutor v. Nikola Šainović et al.,* the ICTY appeals chamber had held "[t]hat 'specific direction' is not an element of aiding and abetting liability under customary international law" (case no. IT-05-87-A, para. 1649, January 23, 2014). Moreover, in *Šainović* the defendant's conviction for aiding and abetting was upheld even in the absence of specific direction.

These precedents were suggestive but not determinative, as our briefs for the opposing side took pains to point out. *Vasiljević* defined aiding and abetting only in passing in a discussion of a different mode of liability, the so-called joint criminal enterprise. As such, the definition was not outcome determinative in *Vasiljević;* that is, it was obiter dictum, in legal terminology. It is widely understood that such passing references are not binding on other courts. In *Šainović,* the rejection of specific direction arguably was determinative for upholding the conviction of the defendant. But *Šainović* did not raise the thorny policy issues of *Perišić* because the defendant in *Šainović* had been part of the same chain of command as the immediate perpetrators and was physically present in the war zone. Hence, *Šainović* arguably concerned a different question and could thus be distinguished—that is, qualified as not relevant—by a competent lawyer. Moreover, it is not clear—and we intentionally did not specify—if precedents are formally binding or merely guiding authority in the ICTY. Finally, the *Šainović* variant of the prosecution's response brief contained an error ("required" instead of "rejected") in its second of three references to the precedent's holding (online appendix Section A1.6, para. 4)

9. Actual group sizes differ because some iPads froze as described in note 10.

(according to their written reasons, however, none of the judges were misled by this).

We made some minor modifications to the precedents to fit them into our case. We changed the date of *Vasiljević* from 2004 to 2014 (that is, after the date of our trial decision) to make it seem natural that the trial court had not referenced the precedent. *Šainović* was decided in 2014 anyway. By contrast, *Šainović*'s discussion of specific direction made extensive references to the *Perišić* decision of (another panel of) the appeals chamber, which implied that there must be another important precedent supporting the opposite outcome. To remove this implication, we changed the text such that the arguments from the *Perišić* appeal judgment referenced in *Šainović* were instead put forward by the defendant in that latter case.

2.4.2. Defendants. The two defendants differed in their nationalities, biographies, and attitudes. We chose these attributes and their depiction to be clearly irrelevant from a strictly legal perspective, at least for the decision at hand. In particular, all the defendants' activities and statements (described below) occurred long after the crimes in question (approximately 2 decades later). They could not reasonably be interpreted to provide any clues about the defendant's mental state at the time of the crime. Even if they did, this would be legally irrelevant because according to the ICTY and our briefs, specific direction is an element of the *actus reus* (the defendant's actions) rather than the mens rea (the defendant's state of mind). Accordingly, our briefs did not make any reference to the defendant's activities and statements in relation to the question of guilt (nor, for that matter, did the written reasons of the participant-judges). Defendants' biographies and attitudes are demonstrably important for lay attributions of blame and perhaps even for sentencing (that is, the length of criminal sentences), but modern legal systems make a strong point of excluding these factors for the determination of guilt in a technical sense (that is, whether a punishable crime was committed in the first place) (compare Nadler 2012; Nadler and McDonnell 2012).

We created two fictitious defendants to avoid the possibility that the judges might recognize a name and be influenced by factors outside of those we provided in the experiment. We named these defendants Borislav Vuković (a fictitious unsympathetic Serb) and Ante Horvat (a fictitious sympathetic Croat).

Vuković's facts and trial judgment were identical to the original *Perišić* facts and trial judgment except for two war crime locations and the name

and some biographical information of the main defendant. As noted above, we omitted all passages relating to Zagreb. We also changed all references to Srebrenica to Vlasenica because we thought that the (real) Srebrenica massacre was too notorious to find a credible Croat equivalent. We made only two substantive, fictitious changes to the *Perišić* original. First, we added the following biographical sentence to the statement of facts and the trial judgment facts: "He held this position [as the army's chief of staff] until his mandatory retirement from the [army] in 2004, when he became advisor to the [Yugoslavian] government for 'the rehabilitation of Serb victims of Albanian persecution' and chairman of the United Serbia Party." Second, the fictitious brief for the prosecution noted in its closing passage that the defendant "has publicly mocked this tribunal and repeatedly inflamed lingering tensions with inflammatory public statements showing absolutely no regrets about the horrors of the war in general, and the war crimes he supported in particular." This statement was supported by a footnote.

Horvat's facts and trial judgment were identical to Vuković's, with the following three exceptions. First, we changed all names of Yugoslav and Bosnian-Serb persons, institutions, and places to their Croat and Bosnian-Croat equivalents. In particular, we changed Vlasenica to Ahmići and Sarajevo (which was besieged by the Bosnian Serbs) to Mostar (which was besieged by the Bosnian Croats). Second, we tried to exploit the fact that Western audiences generally perceived Croatia's role in the war more positively than Serbia's. To reinforce this association, our (fictitious) facts noted in passing that Horvat had contacts with the North Atlantic Treaty Organization (NATO) during the war. Third, we changed the two fictitious bits of information relating to the accused's postwar behavior. After retirement, Horvat "became vice-chairman of the Croatian-Bosnian Reconciliation Commission." The prosecution does not comment on Horvat's attitude. Instead, the defense notes in closing: "From the very beginning of this case, the defendant has expressed his deep regret at all bloodshed in this tragic war, and in particular at the inexcusable crimes of certain soldiers and officers in the field. He categorically denies, however, that he is personally responsible for those crimes. We urge the Appeals Chamber to affirm that the law is on the side of the defendant and others forced by history to make difficult decisions in times of war, and to overturn the conviction by the Trial Chamber."

These descriptions may seem blunt in isolation. It is important to emphasize, however, that they were embedded in longer documents in arguably natural ways. If anything, we were concerned that the judges would overlook these passages.

3. PRIOR BELIEFS: SURVEY OF LAW PROFESSORS' EXPECTATIONS

On the basis of the legal literature and our experience as lawyers, we expected both legally relevant and irrelevant factors to have equally sizeable effects on the judges' decisions. The precedents were weak, but the judges had little else to fall back on and little time to develop their own theories of ICTY law. The differences between the defendants were sizeable but hidden in a few sentences. Concretely, we expected both treatments to shift the affirmance probability by about 40 percentage points. That is, we expected that most judges would overturn Horvat's conviction under the *Vasiljević* precedent and affirm Vuković's conviction under the *Šainović* precedent and that the other two combinations would fall somewhere in the middle. Such effect sizes have been observed in comparable experiments with lay subjects (for example, Norton, Vandello, and Darley 2004; Norton, Sommers, and Brauner 2007).

To verify that our treatments represented interesting variation, however, we surveyed law professors about their expectations for the experiment. We e-mailed a brief description of the experiment to all tenured faculty at four of the top 10 US law schools and to a randomly selected subset at the six others. The e-mail message invited recipients to submit their expectations of effect sizes (none, modest, or strong) via an anonymous link. The response rate was over 25 percent at the first four schools and 17 percent at the latter six. (For the full text of the e-mail survey, see online appendix Section A2.)

In the interest of achieving a decent response rate, our e-mail message's description of the experiment had to be brief. Any brief description inevitably leaves many details to the respondents' imaginations. We hope that the e-mail message's information was sufficient, however, for respondents to have the right idea about the key variations in the experiment whose effects they were asked to predict. In particular, the survey sent to faculty at the six randomly selected schools explicitly listed the key attributes of our defendants, describing them as "a regretful, conciliatory Croat or a hateful, nationalist Serb." (At the four top schools, we described the defendants either as "a likeable Croat or an unsympathetic Serb," with virtually identical results.) These descriptions arguably made the defendants' differences appear more prominent than they really were because the e-mail message did not mention that the defendants' characteristics were discreetly interspersed in a few sentences. As for the precedents, the e-mail message described them as "dicta or distinguishable (because the precedent involved primary perpetrators in the same formal organization as the accused)." This description does not mention that the

Table 1. Law Professors' Expectation of Effect Sizes

Effect of Defendant	Effect of Precedent			
	None	Modest	Strong	Total
None	2	28	9	39
Modest	2	18	29	49
Strong	3	8	3	14
Total	7	54	41	102

Note. The Wilcoxon signed-rank test of defendant versus precedent effect yields $p < 1.4 \times 10^{-8}$; $N = 102$.

dictum in *Vasiljević* was a mere definition without extended discussion and hence rather weak even within the class of obiter dicta. On the other hand, the e-mail message's description also did not point out that the discussion in *Šainović* aimed directly at the prior panel's decision in our case, which made it a rather strong precedent as far as distinguishable precedents go. Overall, our e-mail message's description arguably gave an accurate idea of the distance between the two precedents.

Table 1 summarizes the professors' responses. Most professors expected both the precedent and the defendant to have an effect, notwithstanding the facts that the precedent was weak and the defendants' characteristics were legally irrelevant. The professors also thought, however, that the precedent would have a stronger effect than the defendant's characteristics. A total of 66 professors thought the precedent's effect would be stronger, whereas only 13 thought the opposite. A Wilcoxon signed-rank test overwhelmingly rejects the null hypothesis that the average professor expected the defendant's characteristics to have as strong an effect as the precedent.

4. RESULTS

Table 2 summarizes participant-judges' decisions by treatment condition.[10] For each of the four defendant-precedent combinations, Table 2

10. Cell sizes (that is, the number of participants per precedent-defendant combination) differ by more than one observation because some iPads froze, which interfered with randomization, but there is no reason to think that these failures biased our results. When an iPad froze, the student assistants gave the participant a new one. The randomization mechanism, however, compensated for a frozen iPad only if the participation was explic-

Table 2. Fraction Affirmed

Defendant	*Vasiljević* (Favorable to Defendant)	*Šainović* (Unfavorable to Defendant)	Total
Horvat (sympathetic)	.50 (3/6)	.36 (4/11)	.41 (7/17)
Vuković (unsympathetic)	.75 (6/8)	1.00 (7/7)	.87 (13/15)
Total	.64 (9/14)	.61 (11/18)	.63 (20/32)

Note. Boschloo unconditional exact tests (two sided) yield Vuković − Horvat = .45; $p < .01$; *Šainović* − *Vasiljević* = −.03; $p = 1.00$; (Vuković ∧ *Vasiljević*) − (Horvat ∧ *Šainović*) = .39; $p < .11$; $N = 32$.

itly cancelled, which the freezing generally prevented. Treatment conditions with higher freezing rates would therefore have smaller cell sizes in the ultimate participation groupings (freezing rates were not exactly equal, even though the differences were not statistically significant). Group sizes including frozen participations were 12, 11, 10, and 10 (the two-subject discrepancy between 12 and 10 presumably arose from one explicitly cancelled participation in the 12-member group). It is likely that a participant affected by freezing subsequently received an iPad with a different defendant-precedent treatment combination (we did not have a process for transferring participations). This is not a problem if the freezing occurred before the participant read the instructions because the only materials viewable up to then—the consent form—did not differ by treatment. The instructions differed only in a very minor way, namely, the name of the defendant and the military group he supported (which at that stage would not have meant anything to the participants). It hardly affects the results, however, if we exclude the five participations that began after the first unfinished participation proceeded beyond the instruction stage (approximately 2 minutes into the experiment): the Fisher exact *p*-value for the defendant effect is now .018, and there still is no precedent effect. There is no reason to think that the freezing induced bias. The freezing resulted from the large size of the ICTY judgments included in the materials (trial judgment and precedent). To induce bias, the freezing would have had to be correlated with the potential outcomes in our experiment. For example, iPads would have had to be more likely to fail for participants who judged Horvat's appeal and would have confirmed or those who judged Vuković's appeal and would have reversed. It is theoretically possible that participants with such inclinations were more likely to click on certain documents in a certain order that caused the operating system to choke. We see no indication of this in the code, the click data, or the treatment-specific dropout rates. In particular, an almost equal number of Horvat (five) and Vuković (six) treatments failed. Among the Vuković treatments, an equal number of *Sainović* and *Vasiljević* treatments failed (three each). The only asymmetry is that more Horvat-*Vasiljević* combinations (four) than Horvat-*Sainović* combinations (one) failed, which should if anything have worked against the finding of a defendant effect because *Vasiljević* favored the defendant.

shows the fraction of the judges that upheld the conviction. (Recall that each judge received only one of the two precedents and judged only one of the two defendants.)

Table 2 shows that the precedent made no detectable difference in our sample: the affirmance rates are almost identical. By contrast, the conviction of the unsympathetic defendant (Vuković) was upheld at more than twice the rate of the sympathetic defendant (Horvat). A total of 87 percent of the judges upheld Vuković's conviction, whereas only 41 percent upheld Horvat's. This difference is not only substantively but also statistically very significant. If the two defendants' true affirmance probabilities were equal (the null hypothesis), we would observe such an extreme sampling difference with less than 1 percent probability, even in as small a sample as ours. We calculate this and all other *p*-values using exact methods; that is, we do not rely on large-sample approximations.[11]

We verified that our results are not explained by observable confounding factors, which might fortuitously be present in our sample in spite of randomization. Neither defendant nor precedent differs significantly by the participants' personal characteristics we collected in the exit survey ($N = 30$), namely, age group, gender, professional background (prosecutor or defender), and prior knowledge of international criminal law. The only exception is that all seven female participants were fortuitously assigned defendant Horvat. But controlling for gender does not change our results.[12] Finally, our results are also unchanged if we exclude certain suspicious observations.[13]

11. The Boschloo unconditional exact test we use in Table 2 is the recommended, conceptually superior, uniformly more powerful generalization of the better-known conditional Fisher exact test (Mehrotra, Chan, and Berger 2003). The *p*-values from Fisher exact tests corresponding to those reported in Table 2 are .01, 1.00, and .17, respectively. A standard large-sample *z*-test would yield smaller *p*-values (for example, .008 for the defendant effect), while the Bayesian *p*-value using the standard conservative prior of Howard (1998) is slightly larger at .014. In the robustness checks reported in note 13 and the note to Table 2, we refer for simplicity to the (excessively conservative) Fisher test.

12. We linearly regressed affirmance on defendant, precedent, and controls. The controls were gender alone or in combination with other participant characteristics. We calculated standard ordinary least squares, robust, and bootstrap standard errors. In all permutations, the defendant coefficient is less than −.45 with a *p*-value below 2 percent in all permutations, whereas the precedent point estimate is close to 0 and statistically insignificant.

13. First, our results would be even stronger if we excluded two participants who submitted a judgment but did not formally finish the experiment. One participant cancelled the participation after submitting a judgment, and another did not formally end the experiment with the Finish button. Excluding these two observations reduces the Fisher exact *p*-value for the defendant effect to .007. Second, our results are unaffected by cor-

In view of our small sample size, it is worth emphasizing that our results are not only improbable under the null hypothesis (as summarized by the *p*-value) but constitute strong affirmative evidence for the alternative hypothesis that Horvat would receive (considerably) more lenient treatment. Under conservative ancillary assumptions, our results yield a Bayes factor of 15 for this alternative over the null hypothesis.[14] The Bayes factor is the ratio of the (rational) posterior odds to the prior odds in favor of one hypothesis over another (Jeffreys 1961 [1998]; Kass and Raftery 1995). That is, someone who initially gave a mere 1/16 chance to the possibility that Horvat would be treated more leniently should now be in equipoise, whereas someone who was initially in equipoise should now favor the existence of such an effect 15:1.

Of particular interest is the comparison of precedent and defendant effect sizes. Without imposing any additional structure, only 19 observations are informative for this comparison: those with a precedent favoring the unsympathetic defendant (Vuković-*Vasiljević*) and those with a precedent disfavoring the sympathetic defendant (Horvat-*Sainović*). (In the other observations, both factors work in the same direction and hence cannot be distinguished.) The difference between these two groups' affirmance rates is 75 percent − 36 percent = 39 percent**.** While insignificant

---

recting or excluding judgments on the basis of inconsistent or missing written reasons. Participants' reasons reveal that three of them erroneously acquitted the defendant; that is, the written reasons support affirmance (guilty), but the judgment entered was reversal (not guilty). This affected both nationalities and both precedents, however, and hence correcting these barely changes the results: the Fisher exact *p*-value for the nationality effect is still .018. Similarly, if in addition to this correction we exclude the three judges who did not write down reasons, the Fisher exact *p*-value remains low at .08 or .003, depending on whether we otherwise retain the full sample or also exclude the two participants who did not formally finish. Finally, if we also exclude the one judge who overrode the factual stipulation that the defendant "had rendered practical assistance," we are left with a sample of 21 observations and yet still obtain a highly significant result for the defendant effect ($p = .006$).

14. Calculation of the Bayes factor requires specification of (a distribution over) the details of the competing hypotheses. We conservatively use completely agnostic (flat) prior beliefs. That is, we assume that Vuković's affirmance probability p_V was initially equally likely to be anywhere between 0 and 1 and that Horvat's affirmance probability was p_V under the null hypothesis and equally likely to be anywhere between 0 and p_V under the alternative. We would obtain an even larger Bayes factor if we centered the alternative around our personal prior probability of a 40 percent effect, which was close to the difference we ultimately observed. Inversely, we would obtain a smaller Bayes factor if we center the alternative around a small effect, perhaps on the theory that the effect cannot plausibly be large. Given the large effects observed in related experiments (for example, Norton, Vandello, and Darley 2004; Norton, Sommers, and Brauner 2007), we find this theory to be implausible.

at conventional levels ($p < .11$), this difference is quite informative in the sense that it gives a Bayes factor of 8.7 for the comparison of defendant and precedent effects.[15] To make this more concrete, consider that before the experiment, only 13 of the 102 law professors thought that the defendant would matter more than the precedent. If one treated the distribution of survey answers as a collective prior belief, the posterior odds should thus be $13/(102 - 13) \times 8.7 = 1.27$. That is, after the experiment, the professoriat should view it as more likely than not that the (legally irrelevant) defendant characteristics mattered more for the outcome than the weak precedent.

The judges' written reasons (see online appendix Section A3), however, entirely disregarded defendant characteristics and focused on the precedent. The majority of the judges (20) engaged the precedent. By contrast, only one participant mentioned a defendant characteristic that was not shared by the two defendants (remorse) and dismissed it as legally irrelevant. In general, the judges' written reasons discussed the legal issues in a tone that one would find in real legal decisions (but understandably in less polished language, grammar, and orthography), mentioning both legal and policy considerations but not personal attributes of the defendants.

On the surface, the participants thus followed the standard legal model and did not assign any relevance to factors that legally should not be relevant. But the participants were slightly more likely to mention (73 percent versus 67 percent) and, if they mentioned it, significantly more likely to follow (75 percent versus 33 percent, $N = 20$, one-sided $p = .085$) the precedent when it helped the sympathetic defendant or hurt the unsympathetic defendant.[16] Given the anonymity and zero stakes of the experiment, we think that this selective use of precedent was most likely unconscious. In any event, the discrepancy between professed reasons and actual drivers of decisions is familiar from experiments with lay people (for example, Norton, Vandello, and Darley 2004).

If judges are moved by defendant characteristics but must motivate their decision on the basis of precedent, then a conflict between the two complicates the judges' task. Thus we might expect the judges to need

15. We again use completely agnostic (flat) prior probabilities within each hypothesis as explained in note 14 and consider effects in the predicted direction (that is, lower affirmance rates for Horvat and under *Vasiljević*).

16. The reported p-value is from a Fisher exact test. A standard z-test of equal proportions gives a (one-sided) p-value of .035.

more time with the legal materials if the defendant is sympathetic but the precedent favors affirming the conviction, and vice versa (compare Kennedy 1998). In our sample, judges spent on average about 2 minutes more with the legal materials (precedent and statute) when faced with such a conflict. This difference equals only about 1 standard deviation, however, and is thus not statistically significant.

5. DISCUSSION

5.1. What Legal Models Are Inconsistent with Our Findings?

Our results do not reject the legal model or show that law does not matter. As mentioned before, the precedents were intentionally chosen to be weak, whereas strong precedents might have been determinative. To the extent that the majority of trial cases turn on the facts against the background of strong precedents or otherwise clear law, our results have little immediate application to them (but similar results from fact-finding experiments do; for example, Guthrie, Rachlinski, and Wistrich [2007]).

Nevertheless, the present results are important for understanding the limits of legal reasoning relative to other influences on judicial decision making. A good lawyer should be able to circumvent the weak precedent, as the judges did in our experiment. At the same time, weak precedents are frequently invoked as persuasive authority by judges and other participants in the legal system. Indeed, the participants in our experiment dedicated much of their written reasons to a discussion of the precedent and, to the extent they followed it, tended to portray this as legally required. If weak precedents have no impact on the outcome, as in our experiment, then such attention seems irrelevant, ritualistic, or even misleading, albeit unbeknownst to the judges.

Moreover, the very idea of judging according to the law is that legally irrelevant factors should not influence the decision. This idea is challenged by the strong impact that our description of the defendant seems to have had on participant-judges' decisions. Especially striking is the contrast of the effect of the defendant characteristics, which should not matter, to the absence of a detectable effect of the precedent, which judges treated as if it should matter.

To be sure, one might think that the law must displace nonlegal considerations only when it is sufficiently determinate. But courts, particularly higher courts, routinely decide cases on the basis of law that is as

unclear as it is in our experiment. The case in our experiment was, after all, derived from a real case at the ICTY. Like other courts, the ICTY aspires to be completely impartial, "tak[ing] no side in the conflict" and determining guilt solely on the basis of evidence presented.[17] Our participants were not ICTY judges, but their commitment to this aspiration seems clear from their written reasons and the self-image of the US judiciary. Again, sophisticated observers have long understood that the idea cannot be realized in pure form, but our survey shows that law professors still underestimated the degree of impurity.

5.2. Ecological Validity

We carefully designed our experiment to capture the defining features of real-world judicial decision making. We provided the full setting of a real case including briefs on both sides, a live legal question, and legal materials to consult. We gave the judges time to absorb the materials, reflect, and reach a decision that they had to justify in writing. Last but not least, all the participants were highly professional judges.

The amount of time available (1 hour) and the absence of a hearing correspond to the severe constraints facing many congested courts for many decisions. For example, US district courts on average spend only 1 hour on an entire probation case and resolve many types of motions in 1 hour or less (Lombard and Krafka 2005). The US circuit court judges decide most cases without a hearing and spend an average of only 3 hours per case, including all ancillary or preliminary matters.[18] Practitioners have told us that judges in state courts have even less time per case. In addition, cases that do take more time tend to involve the resolution of many more factual and legal issues than our experimental case. That being said, in many cases judges do have more time per issue, particularly for novel questions of law. The additional time for reflection might reduce the influence of nonlegal factors. On the other hand, the direct im-

17. See International Criminal Tribunal for the Former Yugoslavia, About the ICTY (http://perma.cc/ZN6A-FEWU).

18. In 2014, the US Courts of Appeals decided 34,114 cases on the merits, 80.5 percent of them without a hearing (see Administrative Office of the US Courts, US Courts of Appeals, Judicial Business 2014, table B-10 [http://perma.cc/CE3R-KKS3]). There are 167 authorized judgeships in these courts, which hear cases in panels of three. If these judges dedicated 2,000 hours per year on the cases decided on the merits, they would spend (167 judges $\times$ 2,000 hours $\times$ 60 minutes per hour)/(34,114 cases $\times$ 1 panel per case $\times$ 3 judges per panel) $\approx$ 196 minutes per case. In reality, administrative tasks and cases not resolved on the merits consume some of the judges' time. Judges from other courts sitting by designation may increase the time available per case.

pressions of litigants from a hearing might increase this influence. Moreover, more time may in fact reduce legal constraints because it allows judges an opportunity to craft a reasoning around them (Kennedy 1998).

That being said, our design is clearly not a faithful representation of decision making in any one court, particularly not in the ICTY. Our design casts US judges into an unfamiliar legal system (the ICTY) and mostly an unfamiliar role (as an appeals judge). Our design also has judges decide by themselves without a hearing in an hour, whereas ICTY judges decide in panels after many months of hearings and with the help of judicial clerks.[19] The ecological validity of our design thus presupposes, first, that the characteristic features of judicial decision making transcend the details of individual courts; second, that these characteristics are at least partially shared between trial and appeals courts; and third, that the participants had the skills and time to overcome the lack of familiarity with the particular legal setting. The first presupposition is shared by most of the literature, which treats judicial decision making as a distinct activity. Some of the literature differentiates appeal and trial courts because only the former are focused on questions of law and developing the law (for example, Kennedy 1998). The distinction is hardly rigid, however, as trial courts also treat questions of law, and many courts have dual roles as trial and appeals courts (for example, US district courts hear bankruptcy appeals).[20] Finally, the judges' written reasons demonstrate that the vast majority of them fully comprehended the purposefully straightforward legal question before them and answered it with standard legal arguments. Five judges noted that time was too short or did not submit written reasons, but most arrived at a decision early (the mean time spent with the legal materials before entering judgment was 35 minutes, with negative skew and only five observations above 45 minutes).

The artificiality of the experimental setting should if anything create a bias against our findings. The reason is that the artificiality should make

19. Clerks of course have their own biases (compare Furgeson, Babcock, and Shane 2008a).

20. To the extent that trial judges behave differently when cast in the role of appeals judge in an alien system, it is not clear in which direction their behavior changes. On the one hand, trial judges might feel less constrained by precedent when they cannot be reviewed, their unfamiliar role as appeals judge invites them to develop the law, and the law they apply is not theirs. On the other hand, the precedential value of an appeals decision invites abstraction from the particular litigants before the court, and lack of expertise in a legal system increases the attraction of the guidance offered by precedent, even a weak one. Similarly, the fact that appellate judges usually decide in groups (panels) may attenuate or exacerbate biases (compare Kerr and Tindale 2004).

it easier to block out legally irrelevant but emotionally salient factors and act like a legal automaton. A real-world judgment has a deep impact on the litigants, and often on politically and otherwise important matters far transcending the case at hand. By contrast, in the experiment, nothing is at stake except judges' professional pride. The litigants are known to be fictitious, and information about them had to be slipped in obliquely and could easily be overlooked. The US judges' implicit assumptions about Croats and Serbs, which we were trying to exploit, are presumably not nearly as powerful as those documented for race and other characteristics the judges would face in domestic cases (compare Lane, Kang, and Banaji 2007). (On the other hand, judges may be particularly sensitized to, and thus possibly inoculated against, the biases that are most likely in their court.) Moreover, the judges knew they were being directly observed, which should have made them particularly legalistic. Finally, the judges had been assigned an article about judicial biases (Guthrie, Rachlinski, and Wistrich 2007) before the experiment, which should have raised their awareness of and, presumably, effort to avoid undue influences. From outward appearances, the participant-judges seemed to take the experimental task very seriously.

6. CONCLUSION

In this study, we experimentally investigated the effect of a weak precedent and legally irrelevant defendant characteristics in a setting uniting the key features of real-world judicial decision making. In particular, we had real judges decide a real case in 55 minutes. We found that defendant characteristics appeared to have a strong effect, whereas the precedent had no detectable effect.

In principle, the legal model of judicial decision making can accommodate the irrelevance of the weak precedent. By definition, weak precedents are distinguishable or otherwise circumventable by a legalistic judge. At the same time, weak precedents are in fact frequently cited as support in legal briefs and decisions, including by our participants, and judges make many decisions in circumstances in which no stronger authority than a weak precedent is available. In any event, the strong effect of legally irrelevant defendant characteristics is troublesome for the legal model and particularly striking when compared with the noneffect of the precedent. The deviations from the legal model observed in our experiment are much stronger than law professors had predicted in a survey.

We hesitate to draw policy conclusions until more studies have replicated and refined our results. Taken at face value, our results argue for greater reliance on rules rather than standards (compare Kaplow 1992). In theory, delegating thorny questions to judges may enable a more nuanced evaluation of case-specific facts under accepted general principles and precedents. Our experiment suggests, however, that oftentimes the principles and precedents may be too ambiguous, and irrelevant facts too difficult to ignore, for this delegation strategy to succeed. Alternatively, we need to uncover strategies that reinforce the influence of relevant criteria over irrelevant ones (compare Sood and Darley 2012; Nadler 2012; Sood 2015). In particular, our results would provide a powerful rationale for blinding judges to many details of the facts, for example sanitizing the record before it is sent up to an appeals court.

On a methodological level, we have demonstrated that experimental investigation of long-standing jurisprudential questions under more realistic conditions is possible. But we have also found that the results are mostly consistent with those obtained in less realistic designs, as mentioned in the introduction. If this consistency is confirmed in future studies, then the credibility of the less realistic designs will be greatly enhanced and the need for cumbersome realistic designs concomitantly reduced. This would considerably facilitate further research into the black box of judicial thought processes.

REFERENCES

Braman, Eileen, and Thomas E. Nelson. 2007. Mechanism of Motivated Reasoning? Analogical Perception in Discrimination Disputes. *American Journal of Political Science* 51:940–56.

DeKay, Michael L. 2015. Predecisional Information Distortion and the Self-Fulfilling Prophecy of Early Preferences in Choice. *Current Directions in Psychological Science* 24:405–11.

Englich, Birte, Thomas Mussweiler, and Fritz Strack. 2006. Playing Dice with Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making. *Personality and Social Psychology Bulletin* 32:188–200.

Epstein, Lee, William M. Landes, and Richard A. Posner. 2013. *The Behavior of Federal Judges: A Theoretical and Empirical Study of Rational Choice*. Cambridge, MA: Harvard University Press.

Frank, Jerome. [1930] 2009. *Law and the Modern Mind*. New Brunswick, NJ: Transaction.

Furgeson, Joshua R., Linda Babcock, and Peter M. Shane. 2008a. Behind the

Mask of Method: Political Orientation and Constitutional Interpretive Preferences. *Law and Human Behavior* 32:502–10.

———. 2008b. Do a Law's Policy Implications Affect Beliefs about Its Constitutionality? An Experimental Test. *Law and Human Behavior* 32:219–27.

Greene, Joshua D., Sylvia A. Morelli, Kelly Lowenberg, Leigh E. Nystrom, and Jonathan D. Cohen. 2008. Cognitive Load Selectively Interferes with Utilitarian Moral Judgment. *Cognition* 107:1144–54.

Guthrie, Chris, Jeffrey J. Rachlinski, and Andrew J. Wistrich. 2007. Blinking on the Bench: How Judges Decide Cases. *Cornell Law Review* 93:1–43.

Holyoak, Keith J., and Dan Simon. 1999. Bidirectional Reasoning in Decision Making by Constraint Satisfaction. *Journal of Experimental Psychology: General* 128:3–31.

Howard, J. V. 1998. The 2 × 2 Table: A Discussion from a Bayesian Viewpoint. *Statistical Science* 13:351–67.

Jeffreys, Harold. 1961 [1998]. *Theory of Probability*. 3d ed. New York: Oxford University Press.

Kahan, Dan M. 2010. Culture, Cognition, and Consent: Who Perceives What, and Why, in Acquaintance-Rape Cases. *University of Pennsylvania Law Review* 158:729–813.

———. 2015. Laws of Cognition and the Cognition of Law. *Cognition* 135:56–60.

Kahan, Dan M., David A. Hoffman, Donald Braman, Danieli Evans, and Jeffrey J. Rachlinski. 2012. "They Saw a Protest": Cognitive Illiberalism and the Speech-Conduct Distinction. *Stanford Law Review* 64:851–906.

Kahan, Dan M., David Hoffman, Danieli Evans, Neal Devins, Eugene Lucci, and Katherine Cheng. 2016. "Ideology" or "Situation Sense"? An Experimental Investigation of Motivated Reasoning and Professional Judgment. *University of Pennsylvania Law Review* 164:349–439.

Kaplow, Louis. 1992. Rules versus Standards: An Economic Analysis. *Duke Law Journal* 42:557–629.

Kass, Robert E., and Adrian E. Raftery. 1995. Bayes Factors. *Journal of the American Statistical Association* 90:773–95.

Kennedy, Duncan. 1998. *A Critique of Adjudication* (*Fin de Siècle*). Cambridge, MA: Harvard University Press.

Kerr, Norbert L., and R. Scott Tindale. 2004. Group Performance and Decision Making. *Annual Review of Psychology* 55:623–55.

Lane, Kristin A., Jerry Kang, and Mahzarin R. Banaji. 2007. Implicit Social Cognition and Law. *Annual Review of Law and Social Science* 3:427–51.

Leiter, Brian. 2005. American Legal Realism. Pp. 50–66 in *The Blackwell Guide to Philosophy of Law and Legal Theory,* edited by Martin P. Golding and William A. Edmundson. Malden, MA: Blackwell.

Lerner, Jennifer S., Julie H. Goldberg, and Philip E. Tetlock. 1998. Sober Second

Thought: The Effects of Accountability, Anger, and Authoritarianism on Attributions of Responsibility. *Personality and Social Psychology Bulletin* 24:563–74.

Llewellyn, Karl N. 1940. On Reading and Using the Newer Jurisprudence. *Columbia Law Review* 40:581–614.

Lombard, Patricia, and Carol Krafka. 2005. Final Time Estimates and Adjustment Percentages Used in the Calculation of the Case Weights. Appendix Q in *2003–2004 District Court Case-Weighting Study: Final Report to the Subcommittee on Judicial Statistics of the Committee on Judicial Resources of the Judicial Conference of the United States.* Washington, DC: Federal Judicial Center. http://perma.cc/9XJL-W4GG.

Mehrotra, Devan V., Ivan S. F. Chan, and Roger L. Berger. 2003. A Cautionary Note on Exact Unconditional Inference for a Difference between Two Independent Binomial Proportions. *Biometrics* 59:441–50.

Nadler, Janice. 2012. Blaming as a Social Process: The Influence of Character and Moral Emotion on Blame. *Law and Contemporary Problems* 75:1–31.

Nadler, Janice, and Mary-Hunter McDonnell. 2012. Moral Character, Motive, and the Psychology of Blame. *Cornell Law Review* 97:255–304.

Norton, Michael I., Samuel R. Sommers, and Sara Brauner. 2007. Bias in Jury Selection: Justifying Prohibited Peremptory Challenges. *Journal of Behavioral Decision Making* 20:467–79.

Norton, Michael I., Joseph A. Vandello, and John M. Darley. 2004. Casuistry and Social Category Bias. *Journal of Personality and Social Psychology* 87:817–31.

Paxton, Joseph M., Leo Ungar, and Joshua D. Greene. 2012. Reflection and Reasoning in Moral Judgment. *Cognitive Science* 36:163–77.

Rachlinski, Jeffrey J., Sherry Lynn Johnson, Andrew J. Wistrich, and Chris Guthrie. 2009. Does Unconscious Racial Bias Affect Trial Judges? *Notre Dame Law Review* 84:1195–1246.

Rachlinski, Jeffrey J., Andrew J. Wistrich, and Chris Guthrie. 2013. Altering Attention in Adjudication. *UCLA Law Review* 60:1586–1618.

Russo, J. Edward. 2015. The Predecisional Distortion of Information. Pp. 91–110 in *Neuroeconomics, Judgment, and Decision Making,* edited by Evan A. Wilhelms and Valerie F. Reyna. New York: Psychology Press.

Schwitzgebel, Eric, and Fiery Cushman. 2015. Philosophers' Biased Judgments Persist Despite Training, Expertise, and Reflection. *Cognition* 141:127–37.

Simon, Dan. 1998. A Psychological Model of Judicial Decision Making. *Rutgers Law Journal* 30:1–142.

Simon, Dan, Lien B. Pham, Quang A. Le, and Keith J. Holyoak. 2001. The Emergence of Coherence over the Course of Decision Making. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27:1250–60.

Sood, Avani Mehta. 2013. Motivated Cognition in Legal Judgment—an Analytic Review. *Annual Review of Law and Social Science* 9:307–25.

———. 2015. Cognitive Cleansing: Experimental Psychology and the Exclusionary Rule. *Georgetown Law Journal* 103:1543–1608.

Sood, Avani Mehta, and John M. Darley. 2012. The Plasticity of Harm in the Service of Criminalization Goals. *California Law Review* 100:1313–58.

Spellman, Barbara A. 2010. Judges, Expertise, and Analogy. Pp. 149–64 in *The Psychology of Judicial Decision Making,* edited by David Klein and Gregory Mitchell. New York: Oxford University Press.

Spellman, Barbara A., and Frederick Schauer. 2012. Legal Reasoning. Pp. 719–35 in *The Oxford Handbook of Thinking and Reasoning,* edited by Keith J. Holyoak and Robert G. Morrison. 2d ed. New York: Oxford University Press.

Wistrich, Andrew J., Jeffrey J. Rachlinski, and Chris Guthrie. 2014. Heart versus Head: Do Judges Follow ethe Law or Follow Their Feelings? *Texas Law Review* 93:855–923.