ISSN 1936-5349 (print)
ISSN 1936-5357 (online)

HARVARD

JOHN M. OLIN CENTER FOR LAW, ECONOMICS, AND BUSINESS

EXPERIMENTAL INVESTIGATIONS OF JUDICIAL DECISION-MAKING

Lukas Holste
Holger Spamann

Discussion Paper No. 1096

03/2023

Harvard Law School
Cambridge, MA 02138

Experimental Investigations of Judicial Decision-Making

Lukas Holste* & Holger Spamann**

Abstract: We review experimental research on judicial decision-making with a focus on methodological issues. First, we argue that only experiments with relatively high realism, in particular real judges as study subjects, plausibly generalize to judicial decision-making in the real world. Most experimental evidence shows lay subjects to behave very differently from expert judges in specifically legal tasks. Second, we argue that studying the effects of non-law is not a substitute for studying the effects of law since large unexplained residuals could be attributed to either. Direct experimental studies of the law effect are few and find it to be puzzlingly weak. Third, we review the substantive findings of experiments with judges, distinguishing between studies investigating legal and non-legal factors and paying close attention to the nature of the experimental task.

Keywords: experiments, judicial decision-making, expertise, effect of law, methodology

* LL.M. (2022), Harvard Law School. lholste@llm22.law.harvard.edu.
** Lawrence R. Grove Professor, Harvard Law School. hspamann@law.harvard.edu.
Thanks to Kevin Tobia and Jeffrey Rachlinski for very helpful comments.

I. INTRODUCTION

We review experimental research on judicial decision-making, by which we mean decision-making by judges in their professional capacity.[1] Which factors influence judicial reasoning and decisions is a central question in legal scholarship. It has obvious implications for law's legitimacy, jurisprudential theories, and comparative law. A burgeoning literature tackles the question with theory, anecdote, and observational data. Experiments promise to isolate causal factors much more cleanly.

Our first two sections address methodological issues. Part II (Methods I) argues that only experiments with relatively high realism, in particular real judges as study subjects, plausibly generalize to judicial decision-making in the real world (so-called ecological validity). Real judges are very hard to recruit for experiments, but lay subjects seem to behave very differently from expert judges in specifically legal tasks. Part III (Methods II) attends to difficulties testing the effect of law.

Part IV reviews the substantive findings of the experimental literature. In brief, judges are strangely unmoved by the law, and subject to the same heuristics and biases as lay subjects, if perhaps somewhat less so. Consistent with our methodological discussion, we distinguish between studies investigating legal and non-legal factors, focus exclusively on experiments with judges rather than laypeople, and pay close attention to the nature of the experimental task. For a complementary survey that treats judge and lay studies equally and integrates basic and applied experiments, see Engel (2022).[2] We limit our review to controlled experiments since natural experiments—real-world settings where a critical feature of interest happens to be quasi-randomly assigned—present different methodological issues.[3]

II. METHODS I: REALISM AND ECOLOGICAL VALIDITY

The challenge is to design lab experiments that have ecological validity for judicial decision-making in the real world. Experiments isolate causal factors within the experimental setting (internal validity). However, we do not do experiments for their own sake; we want them to tell us something about the real world (ecological validity). Field experiments have ecological validity by definition but are impossible without collaboration from legal authorities that thus far reject the idea.[4] This leaves lab experiments.

---

[1] Our definition excludes much valuable experimental work in and around law but not involving judicial decision-making. For example, MacLeod (2019, 2021) and Tobia et al. (2022) test the "ordinary meaning" of certain terms, which many judges purport to refer to in their reasoning, but which by definition is not specific to judges.

[2] Prior reviews include Braman (2017) and Rachlinski & Wistrich (2017). For reviews of the non-experimental empirical literature on judging, also see Harris & Sen (2019).

[3] Whereas lab experiments' weak spot is ecological validity, natural experiments' weak spot is treatment assignment: whether it was truly (quasi-)random (i.e., orthogonal to the effect of interest), and how to adjust statistical inference when treatment is not independently assigned at the individual level but, e.g., at the level of courthouses (e.g., Engel & Weinshall 2020). A problem particular to natural experiments of judicial decision-making is case selection, i.e., litigants may strategically change the cases they litigate in response to the otherwise quasi-random variation. For example, one typical natural experiment exploits discontinuous regime change along a continuous dimension such as time or space (e.g., Hofer 2007; Yang 2015). For legal changes, however, litigants and potential litigants may change the cases they will bring on either side of the threshold, and perhaps other behavior as well (e.g., Hubbard 2017).

[4] Engel (2022) worries about the ethicality of field experiments in the judicial realm. At least some field experiments of judging would not raise ethical concerns. For conceptual discussion of and advocacy for field experiments in courts, see Green & Thorley (2014); Kopas & Thorley (2018); Lynch et al. (2020).

This section reviews theory and empirics suggesting that ecologically valid lab experiments of judicial decision-making must exhibit some degree of realism, above all the participation of real judges. Judicial decision-making involves specialized tasks performed by judges who are expert in these tasks. One day, the basic science of human decision-making may progress to the point that we can design simplified experiments isolating the specificities of the judicial task even with lay subjects, and predict real-world judicial behavior from such experiments.[5] Engel (2022) approaches the existing literature from this "basic science" perspective. By contrast, we take an "engineering" perspective because the gap between judicial decision-making and the current state of the basic science of human behavior appears to us too large. As we report below, major differences in behavior on judicial tasks have been documented between judges and lay people. Known psychological theories, including those surveyed in Engel (2022), do not explain these differences or predict how judges' real-world behavior will differ from students' behavior in the lab. Until we have such theories, we need applied experiments much like engineers and doctors.

A. Subjects

Judicial decision-making has many components that are candidates for judicial expertise in the technical sense of reliably superior performance (Spellman 2010; see generally Ericsson 2018). Such expertise would be limited to a very narrow domain, possibly specific to a particular type of judge.[6] Most importantly, judicial decisions are supposed to be guided, even determined, by decision criteria laid down in a complex edifice of ideas and materials, namely the law. Judges have years or even decades of training and experience in legal analysis and are selected specifically for this skill. This may give them an extraordinary ability to apprehend a legal situation and to interpret and apply the legal materials (Llewellyn 1930; Schauer 2010; Kahan et al. 2016; Spamann & Klöhn 2016)—or to manipulate them, for that matter (Kennedy 1998).[7] Other seasoned legal practitioners might approximate judges' legal expertise but MTurkers, college students, or even law students would not.

Even for decisions where judges are not experts in the sense of delivering *superior* performance, they may develop a routine that generates *different* performance. Practice alone does not necessarily make perfect.[8] Still, it would be odd to expect a law student confronted with a

[5] See, e.g., Zeiler (2010) for an argument along these lines in other legal applications (not judicial decision-making). For a review of behavioral theories of judicial decision-making, see Teichman & Zamir (2014); cf. Simon (2019) (applying psychological theories to jury decision-making).

[6] Relatedly, differences in training and experience may mean that judges in some jurisdictions systematically differ from lay people for a particular task while judges in other jurisdictions do not. Empirically, Spamann et al. (2021) find significant differences in thought processes and written reasons between judges from the U.S., Argentina, Brazil, China, France, Germany, and India.

[7] *Cf.* Adam Liptak, *An Exit Interview with Richard Posner, Judicial Provocateur*, available at https://www.nytimes.com/2017/09/11/us/politics/judge-richard-posner-retirement.html (https://perma.cc/2NP4-B6JM) (quoting Judge Posner: "When you have a Supreme Court case or something similar, they're often extremely easy to get around."). In psychological terms, judges' expertise might be limited to system 2 (deliberative decision-making) once activated, or it might consist in activating system 2 to override system 1 (intuitive decision-making), or it might (perversely?) allow their system 2 to generate reasons to validate system 1.

[8] An additional requirement for expertise to develop on many tasks is instant, frequent, and unbiased feedback (Kahneman & Klein 2009). Judges arguably get such feedback for many of their legal pronouncements from professors and colleagues, and often got such feedback even before joining the bench, for example when making legal arguments to other judges and counsel as a litigator. But often judges do not get such feedback for the many complex factual determinations they have to make, in particular for such high-stakes yet high-volume decisions as setting bail or

decision for the first time to decide even approximately like a judge who has done it thousands of times. The bottom line is that the more specialized the task, the less one can assume that judges behave like other experimental subjects.

Judges likely differ from others also by selection and socialization (beyond their possible effects on expertise). Not everyone with the skills and experience, let alone everyone with the raw talent, to become a judge chooses and is chosen to become one. Once on the bench, judges are subject to social expectations that may shape what they want to do, independently of what they can do. Anecdotally, judges are extraordinarily conscientious in experiments in our and others' experience; Rachlinski et al. (2008, p. 1244) systematically document one such instance.

Experiments comparing judges and other participants largely bear out these concerns and conjectures.[9]

We find big differences in the area where judges have the biggest a priori edge: legal reasoning. Spamann & Klöhn (2023) explicitly test the difference between U.S. federal judges and law students in the relatively high-realism task of Spamann & Klöhn (2016): the decision of an appeals case with briefs, legal materials, and almost an hour of time. In their 2×2 experiment cross-varying a precedent and defendant sympathies, students are only moved by the former and judges only by the latter, and equality of the two effects between the two groups is strongly rejected. Spamann & Klöhn (2023) also find that students significantly differ from judges in the reasons they write and in (the observable trace of) their reasoning process: their view path of the legal documents in the experiment. Turning to vignette studies, Redding & Reppucci (1999) found that sociopolitical attitudes inflected judgments of legal relevance and admissibility of evidence—as described in vignettes taken from judicial opinions—only in students but not judges (both from the U.S.). Kahan et al. (2016) asked U.S. participants in two separate vignettes whether a vignette-specific statute applied, randomly changing legally irrelevant but culturally sensitive facts such that each participant judged one "conservative" and one "liberal" behavior. They found predictable differences in interpretation in function of participants' cultural worldview for lay participants in both vignettes, for law students in one, and for lawyers and judges in neither.[10] That said, Chinese judges and law students were similarly influenced by mention of another court's decision in the vignette study of Chen & Li (2018).[11]

---

deciding on asylum. These hinge on determinations of flight and re-offense risk and the risk of persecution, respectively, neither of which the judge subsequently observes, at least not in a regular, unbiased way.

[9] Our discussion in the main text omits studies that do not cleanly address the question, usually because that was not their goal. Some studies find differences between judges and other participants but do not hold the experimental design constant between the groups, especially their roles, thus introducing a confound: Bushway et al. (2014, n. 10) (judges, defense attorneys, and prosecutors cast in their respective roles); Struchiner et al. (2020); Rachlinski et al. (2013a, at 1210) ("similar materials and a nearly identical apology script" as Robbennolt (2003) but different roles [judge vs. party] and accident). Inversely, Chorn & Kovera (2019) find no differences in "judge, attorney, and mock juror decisions about psychological expert evidence" but perhaps would have found them had they not cast the participants in their respective roles. Landsman & Rakos (1994) report "results suggest[ing] that judges and jurors may be similarly influenced by … biasing material," but the task and hence bias were ill-defined because subjects were asked to adjudge liability even though they had not seen *any* evidence.

[10] See figures 5 (p. 394) and 6 (p. 396) in Kahan et al. (2016). Kahan et al. does not present formal statistical tests for the difference between students' and judges' effects. In its corresponding regressions tables A1 (p. 425) and A2 (p. 429), the difference between the appropriate combination of interaction terms (of worldview interacted with case facts) is relatively small and probably not statistically significant in the first but larger and probably statistically significant in the second.

[11] Chen & Li's finding of a precedent effect for Chinese judges mirrors that of the higher realism study of Liu et al. (2021).

In fact-finding and evaluation—often left to juries—many but not all studies find judges to have an edge too (see generally Bornstein & Greene 2017, 275-82).[12] In Norway, Wessel et al. (2006) find judges but not laypeople to be able to ignore a (mock) witness's emotional expression. Hastie & Viscusi (1998) and Viscusi (2001) find U.S. judges to be better at risk assessment (inter alia, less hindsight bias—cf. Rachlinski et al. 2011) and conversion to negligence rulings than lay people. McQuiston-Surrett & Saks (2009) find that the form of presentation of expert witness testimony influences the weight given in a determination of guilt only by jurors, not state judges. Tobia (2020) finds intentionality ascriptions to shift gradually from U.S. laypeople to law students to judges. By contrast, Lindholm (2008) finds Swedish judges to be no better than laypeople at identifying false testimony. Rosenblatt et al. (1989) find both judges and college students to set higher bail for a prostitute when their mortality is made salient. Hornuf & Klöhn (2019) simply find students to be more confused, but their tasks are even less realistic than those in Wessel et al. (2006) and Lindholm (2008). Finally, Miller (2019) finds judges to be *more* gender-biased than lay subjects in vignette decisions of a child custody and an employment discrimination case.

In tasks that are not specific to judges' expertise, judges exhibit the same behavior as laypeople.[13] For example, judges display biases on the Implicit Association Test (IAT) (Levinson et al. 2017) and perform slightly worse than Harvard college students on the Cognitive Reflection Test (CRT) (Guthrie et al. 2007; 2009; also see Helm et al. 2016 [arbitrators]). Unsurprisingly, slapping legal labels onto the tasks does not change this. For example, Sonnemans & Van Dijk (2012) find that Dutch candidate judges and law students perform about the same on two abstract decision-theoretical tasks, notwithstanding that the inference goal was labelled determination of guilt. Similarly, Pantazi et al. (2020) find that judges and laypeople have equal difficulty ignoring bold-face false statements interspersed into a normal-type vignette—a task with no correspondence in real-world judging, or any other real-world activity for that matter—even though the vignette described facts of a potential crime and the task was to determine guilt. These comparative findings are consistent with a prominent line of research with judges alone that shows them to be mostly prone to the same heuristics and biases as other human beings (see infra 3.2). Judges are human, and surely this has repercussions for their professional behavior. Still, they seem to perform differently within their narrow expert domain.

One major exception where judges react like laypeople in a specialized task is anchoring bias in sentencing. Sentencing is an archetypical judicial task. Nevertheless, judges exhibit similar anchoring bias as laypeople even with relatively realistic materials (albeit not a realistic task[14])

[12] The non-experimental literature comparing judges and juries—necessarily limited to fact-finding tasks—finds that they diverge in about one fifth of the cases in predictable ways (Kalven & Zeisel 1966; Vidmar & Hans 2007, 148-151). Wissler et al. (1999) and Robbennolt (2002) compare experimental damages awards by judges and jury-eligible citizens. See Robbennolt (2005) for a survey of the literature comparing judges to juries.

[13] In addition to the studies discussed in the main text, Redding & Reppucci (1999) and Schmittat & Englich (2016) find that expert judges, non-expert judges, and students are equally unlikely to ascribe importance to arguments—irrespective of their legal admissibility—against their position after they have made up their mind. Both studies frame this as a bias. But as Kahan et al. (2016, at 367-8) point out with respect to Redding & Reppucci, there is no bias in "sticking to one's guns" if the information content is low relative to the information one already has and/or the factual information is irrelevant given one's normative commitments. Schmittat & Englich claim that judges differ by expertise but they drop one of two outcome measures for lack of variation and their statistical analysis finds "no main effects of Expertise" for the other (at 394).

[14] The task in Englich & Mussweiler (2001) and the Englich et al. studies is not very realistic because the judges there sentence defendants without ever having seen them. In practice, sentencing happens in person. Moreover, in most of the world (but not the U.S.), sentencing is inevitably preceded by an in-person trial presided over by the same judge.

(Englich & Mussweiler 2001; Englich et al. 2005, 2006).[15] Perhaps anchoring is a particularly universal bias. Alternatively, sentencing may be a particularly vulnerable task. Sentencing explicitly grants judges discretion in a way that legal interpretation does not.

In summary, researchers need to think carefully about the appropriate study population for their research question. If the question involves specialized judicial tasks, only an experiment with judges will generate answers that can plausibly generalize to the real world. This is unfortunate because recruiting judges as experimental subjects is difficult. Judges are few and busy. Organizations that could mobilize them in greater number—courts, judges' associations, and judicial agencies like the FJC—are wary of doing so.[16] Using other study subjects, such as law students or laypeople, would allow more experiments with larger sample sizes at lower cost. For non-specialized tasks, that is the right way to go. For specialized tasks, however, we see enough evidence of differences between study populations to distrust investigations of specialized judicial decision-making using lay or student subjects. Legal professionals are more promising as stand-ins for judges, but whether they really are would be a highly valuable subject of future research.

B. Tasks

Judges do not usually make quick decisions in a vacuum. As already mentioned, a characteristic feature of judging is that it is supposed to be guided, even determined, by an elaborate edifice of norms. These norms are applied to facts that are often painstakingly fleshed out in protracted proceedings. These proceedings follow a decorum—wigs, robes, benches, gavels, flags, and all—that is supposed to underline the solemnity of the task and the necessity for justice to be blind and even-handed. To showcase and safeguard the impartiality and quality of the decisions, hearings are public and judges must usually give reasons, usually in writing; there is also often the possibility of appeal to a higher court. All of this is done because of the weighty consequences that attach to real-world judicial decisions. A judge might well be perfectly law-abiding, measured, and impartial in such setting even while acting in an unprincipled, rash, and biased manner in an experiment.

Short of a field experiment, no experiment can fully simulate this environment. Even in an imaginary experiment with elaborate role play over several weeks in a mock courtroom, participating judges would still know that they are in an experiment and that their decisions will not have any direct consequences for litigants or the legal system (setting precedent) (e.g., Kihlstrom 2021). Nor would they be subject to the same transparency and oversight.[17]

---

[15] Also see for judicial anchoring in less realistic materials, e.g., Guthrie et al. (2001), Wistrich et al. (2005), Rachlinski et al. (2006), Guthrie et al. (2009), Rachlinski et al. (2015) (summarizing prior studies at 707-709); Spamann et al. (2021) (whose anchoring part was a minor add-on to the much more realistic main study); and for arbitrators Franck et al. (2017).

Bystranowski et al.'s (2021) meta-study of anchoring in legal experiments finds only slightly and not statistically significantly less anchoring in studies with "professionals" than with lay subjects. The "professionals" category contains 11% law students. The study distinguishes civil and criminal law tasks but not whether the task involves judicial discretion.

[16] This could be because the expected results are uninteresting to them, or because they are afraid of what the results would show.

[17] Making participants' experimental decisions and opinions public and/or relevant to their careers is unlikely to be palatable to judges and the organizations that can provide access to them. Worth trying would be an intermediate design that reveals participants' decisions and opinions only internally to other participating judges—perhaps sitting in "appeal" over initial participants' decisions.

Ecological validity thus hinges on a plausibility conjecture specific to the task and the effect of interest: does the experimental task plausibly represent those features of the real-world judicial task that might most likely moderate the effect of interest? We can never test such a conjecture because we have no field experiment for comparison. Nonetheless, it would be foolish to treat individual snap judgments in short vignette studies as evidence of the behavior of collegial courts in extended high-stakes proceedings such as the U.S. Supreme Court. By contrast, longer, more elaborate experiments may well approximate many of the more prosaic decisions that most judges make most of the time.[18]

It seems critical that the experimental task is of a type that judges actually encounter in real life (e.g., judging a legal question on motions or appeal[19]). This may include litigants' briefs to frame the case and legal materials to provide the applicable decision rules. It arguably requires a task-appropriate amount of time and, for most decisions, reason-giving (usually in writing) because judges are required to do so in real life as well. Liu (2018) experimentally confirms the importance of giving reasons.

At the absolute minimum, tasks purporting to test legal reasoning need to be labelled as such. Tobia (2020) finds that U.S. judges interpret "intentional" differently depending on whether the question is asked in the abstract or in the context of whether an "intentional" crime has been committed.

Extrapolation may be possible even if some judicial behavior cannot be captured well in an experiment (e.g., the U.S. Supreme Court's). For example, if judges do not follow pertinent precedent even in relatively simple, anodyne cases, then it is hard to believe they would actually be guided by the many marginally relevant precedents they tend to cite in long opinions in difficult, controversial cases, at least when appeal is not possible (Spamann et al. 2021; Klerman & Spamann 2024). In general, however, extrapolation is difficult to impossible because we do not know in which direction a lack of realism would bias results. For example, are judges in experiments likely to be less legalistic because ascertaining the law is hard and they do not have the pressures of appeal and reputation, or more legalistic because there is no temptation to bend the result towards a desired result for the litigants? Similarly, do richer materials make judges less biased because they emphasize the judicial role and may contain more detailed constraints, or more biased because the more materials there are, the more experts can manipulate them to reach a desired result? Ideally, such questions would themselves be studied experimentally.

III. METHODS II: STUDYING THE EFFECT OF LAW

Studying the effect of law poses special challenges. Most empirical studies of judicial behavior vary and hence study the effect of non-law, not law. This includes all studies of judicial bias. Nonetheless, law is the most distinctive and central aspect of judicial decision-making.

Studying the effects of non-law is not a substitute for studying the effects of law. To be sure, to the extent a decision is driven by specific non-law, it is not driven by law. The reverse, however, is not true. The absence of some bias does not mean that decisions are driven by law—there might

[18] Anecdotally, participants in the relatively high realism studies of Spamann & Klöhn (2016), Spamann et al. (2021), and Klerman & Spamann (2024) approached the task with the utmost seriousness in the venues that one of us (HS) was able to observe, and their written reasons mostly adopt a distinctively judicial tone.

[19] By default, most U.S. judges do not adjudicate facts at trial; that is reserved to juries. When judges (or anyone else, for that matter) do adjudicate facts, they do so on the basis of evidence. An experiment on judicial fact-finding should thus construct a setting where judges do indeed find facts, and where the evidence is of a type that can be presented in an experiment.

be other biases or simple randomness (i.e., noise) at work (cf. Fischman 2014). The point is quantitatively important. While many studies have found non-legal influences on judging, the estimated magnitudes tend to be small. If one erroneously interpreted as legally determined anything that is not driven by a documented bias, one would conclude that the law effect dwarfs non-law effects in judging. By contrast, experimental studies of the law effect find it to be puzzlingly weak (*infra* IV.A).

Varying the law is hard, especially if one aims for a high degree of realism (*supra* Part II). The law that exists is set, and judges know or at least may know it in jurisdictions they work with in real life. Selectively revealing legal authorities (as in Spamann & Klöhn 2016; Liu et al. 2021; Spamann et al. 2021) works only if the judges do not know the existing authorities, which requires use of law they are not intimately familiar with, which may come at a cost in terms of realism.

Probably the best option is varying the date or location of the case or proceedings in a way that changes the applicable law. Even this option has pitfalls. The location and hence applicable law must not be unrealistically alien. Date and location must not have plausible relevance for the judge and situation in question other than via legal commands.[20]

Varying other facts relevant under the law is not an alternative. The reason is that almost all legally relevant factual variation may also make a difference for non-legal reasons.[21] In particular, legally relevant facts almost always have normative valence. Judges almost surely have normative preferences. Thus, if judges react to a legally relevant fact, they may be reacting because of their normative preferences, not because of anything contained in the law.[22] For example, if Ontarian judges tend to view intoxication as an aggravating factor in sentencing (Macdonald et al. 1999) while French and German judges ignore it (Bègue et al. 2020), this may be because Ontarian law differs from French and German on this point or because Canadian and European judges have different conceptions of criminal justice in such cases. This distinction might not be of practical import to a criminal defendant or prosecutor, but it would be for a legislator trying to control judicial behavior. Date and location are different only because and to the extent they do not have normative valence and matter only as arbitrary lines in legal rules.

Empirical researchers may wish to sidestep longstanding debates on what counts as law. Factors like non-binding sentencing recommendations (Bushway et al. 2012—a natural experiment) or ministerial guidelines (Bourreau-Dubois et al. 2021) are of obvious interest for legal practice and theory and thus a worthy object of study irrespective of whether they are technically "law". If an incontrovertible test of "law" is desired, however, it is preferable to vary instruments that are "law" by any definition (as in Klerman & Spamann 2024).

In most imaginable experiments testing law effects, a skilled lawyer could formulate an argument why a legal variation that did not have an effect in the experiment should not have had an effect because the content of the law, *properly understood*, was unchanged; and vice versa.[23]

---

[20] This condition is arguably violated for study of choice-of-law rules themselves, i.e., meta-rules that tell the judge which state's substantive law to apply. Judges might well have normative *preferences* for choice of law, confounding the possible effects of a choice-of-law *rule* when changing location facts. See Klerman & Spamann (2024).

[21] Similar problems have long plagued observational studies. See Klein (2017), at 242.

[22] To be sure, if the judges do *not* react to a factor that should change decisions under the law, then one can infer that they do *not* follow the law, and if they do react, one can at least infer that the fact has a causal effect *consistent* with the law. But one cannot interpret the presence of an effect as a causal effect *of law*—a difference that would very much matter if one were to, e.g., change the law.

[23] Similarly, a skilled lawyer can often construct an argument why some seemingly extra-legal factual variation should matter under the law after all. The argument in the main text applies analogously.

Sometimes the argument would surely be justified; it would mean that the experiment was badly designed. However, when the legal implication of the legal variation seems clear to most observers and the *possibility* of an argument against this implication is raised to defend "the law" against the—experimentally supported—charge that it has no effect, then this defense is ultimately self-defeating. It defuses the charge of ineffectiveness only by substituting an inordinate degree of legal indeterminacy, which most people would view as similarly problematic (cf. Fischman 2014 for the case of inter-judge disparities).

IV. LITERATURE REVIEW

This section summarizes experiments to date. For reasons given in section II.A, we focus on experiments with real judges. All experiments are lab experiments, with the exception of three field experiments that we will explicitly label as such.[24] While we have endeavored to include every major study not yet mentioned in the prior parts, we make no claim to comprehensiveness given the volume of the literature. We organize our discussion by the factors under study: law, non-law, and factors that might be either.

We omit as outside our topic an earlier literature simulating judicial behavior without experimental (i.e., randomized) variation.[25] That said, non-randomized aspects of an experimental study can produce useful information for non-causal questions. Spamann et al. (2021) track judge-participants' document use in their experiment, which they administer in seven jurisdictions. They find systematic differences in this outward manifestation of judicial thought between the seven jurisdictions, but not between those belonging to the civil law on the one side and those belonging to common law on the other, thus refuting the idea that common and civil lawyers think fundamentally differently, at least at this level of measurement.

An important generic caveat is that most of the studies reported here do not yet embrace research practices such as preregistration and disclosure of methods, data, and materials that are increasingly considered indispensable for credibility in science (cf. generally Nelson, Simmons & Simonsohn 2018). As in other lines of research, it is likely that results are much more likely to get reported if they show an "interesting" effect, especially bias (cf. Franco et al. 2014; Spamann 2022). Similarly, some studies testing, e.g., multiple biases highlight only those tests that rejected the null hypothesis, which may lead to the perception that bias is more widespread than it actually is. Preregistered replications of these experiments would be highly valuable.

A. Experiments Testing the Effect of Law

Despite law's centrality for judicial decision-making (*supra* Part III), there are very few studies of the effect of law. Those that exist find a surprisingly small actual effect of law, if any. It remains to be seen if these results replicate. Even at face value, however, these findings are not inconsistent with a high degree of consistency in real-world judicial decision-making because such consistency can be created by consistent judicial preferences and judicial hierarchy even if legal authorities (precedents, statutes) have no guiding force per se.

The main existing studies of law effects are relatively high realism studies that we will refer to as "the ICTY experiment" (Spamann & Klöhn 2016; Liu et al. 2021; Spamann et al. 2021) and

[24] There are, of course, field experiments varying support, and tracking outcomes, for litigants etc., which we consider outside our topic.

[25] Such studies include Becker (1966), Hood (1972), Lemon (1974), Austin & Williams (1977), Palys & Divorski (1986), Doob & Beaulier (1992).

"KS" (Klerman & Spamann 2024, which was preregistered). Both gave real judges an hour to decide a fictionalized case presented with a full set of legal materials—briefs, facts and/or trial judgment, legal authorities (statute, precedent, etc.). Judges were to give brief written reasons. The posture of the case was such that the judges had to make a decision only on the law, not facts. To run and compare the study in seven jurisdictions[26], the ICTY experiment employed an international criminal case, which is not realistic for its domestic judge participants. KS remedied this shortcoming, employing a purely domestic U.S. civil case administered to U.S. judges.

These high realism studies find a surprisingly small law effect, if any. Both the ICTY experiment and KS randomly assigned one legal factor and one non-legal factor. The legal factor was a horizontal precedent in the ICTY experiment and the forum's choice of law principles in KS. The non-legal factor was sympathies of the parties in both experiments. In the ICTY experiment, rates of affirmance of defendant's conviction—the decision to be made by participants—differed only by a few percentage points between radically different precedents, less than between defendants.[27] It could be argued that the horizontal precedents in the ICTY experiment were not legally binding, such that judges were at liberty to disregard them, however pertinent their holdings. In KS, however, the legal factor was indisputably binding. Nonetheless, KS's point estimate of law following is only 11% above a fair coin flip, i.e., judges followed the law 61% of the time when only two choices were available (more so under a rule than a standard); even the upper 95% confidence interval is only 26% above the coin flip. That only the ICTY experiment found a sympathy effect might be because the facts in KS are much less wrenching. Alternatively, the ICTY experiment may have triggered an in-group bias that has also been documented in natural experiments of Israeli bail decisions (Gazal-Ayal & Sulitzeanu-Kenan 2010) and Kenyan criminal appeals (Choi et al. 2022); KS's sympathy manipulation was arguably less prone to triggering this.

A corollary of small law effects is that inter-judge disparities can be very large. Bourreau-Dubois et al. (2021) report that non-binding ministerial guidelines reduce the variance of child support awards in a vignette experiment with French judge-apprentices. Since the reported reduction is only about one fifth of the overall variance, however, it might be more appropriate to emphasize that high inter-judge variability persists *in spite of* the guidelines.[28] This is consistent with observational data showing massive inter-judge disparities even with randomly assigned cases (e.g., City Magistrates' Courts, City of New York 1917, pp. 35, 73; Ramji-Nogales et al. 2007; Fischman 2014; also see Engel & Weinshall 2020 for variation induced by caseload).

How can the absence of large law effects be reconciled with the judicial practice of purporting to provide legal reasons for their decisions? Liu (2018) and Liu & Li (2019) investigate this question in vignette studies with Chinese judges. Liu (2018) varies not only the requirement but also the timing of reason-giving. This can be interpreted as a variation of *procedural* law (albeit not the kind of variation we usually see in the real world). In real life, judges generally write reasons after deciding, and/or delegate writing to clerks. Liu finds that writing down reasons—or at least being forced to think—*before* deciding eliminates bias, whereas delegating the writing of

[26] Argentina, Brazil, China, France, Germany, India, and the United States.

[27] Sample sizes for individual jurisdictions in the ICTY experiment are too small for reliable effect size comparisons. For what it is worth, however, the ICTY experiment found the least effect of precedent in the United States (Spamann & Klöhn 2016) and the greatest in China (Liu et al. 2021) (see generally Spamann et al. 2021); the latter finding is consistent with the vignette study of Chinese judges of Chen & Li (2018), which varied information about the decision of a sister court.

[28] In fact, the reduction may be nil because the study randomizes at the level of groups, not individuals, and does not statistically account for possible pre-treatment differences between the groups.

reasons to clerks exacerbates it. Liu & Li (2019) take a closer look at judges' reasons themselves (without varying the law). Their Chinese judge-participants are influenced by a legally irrelevant factor (i.e., they are biased) but do not mention it in free-form reasons and a multiple-choice questionnaire. This is consistent with the behavior of U.S. judge-participants in the ICTY experiment (Spamann & Klöhn 2016).

At the intersection of judges' written reasons and non-law sits the field experiment of Thompson et al. (2022). They create Wikipedia articles for random highly-cited Irish Supreme Court cases and find that this increases judicial citations to these cases by over 20%. Like other elements of judges' written reasons, citations may be unrelated to decisions, as we just argued. Moreover, changing Wikipedia is definitely not changing the law. Nonetheless, the experiment speaks to judges' perception, representation, and perhaps development (by citations!) of the law.

B. Experiments Testing the Effect of Non-Law

The bulk of the experimental literature on judicial decision-making tests the effect of non-law. Most of these studies interpret their findings as evidence of bias, implying that the factors they study should not matter under the law, i.e., that they are "extra-legal factors." As we occasionally remark, however, this interpretation may sometimes be unwarranted.

We are aware of only a single field experiment. Krasno et al. (2021) send elected judges in Wisconsin letters demanding recusal in particular cases, pointing out that one of the attorneys appearing in the case donated to the judge's election campaign. This increases the rate at which the judges disclose the conflict but not their rate of recusal.

Lab experiments could be further divided by whether they study a purely legal task, a purely non-legal task, or something in between. Besides the ICTY experiment and KS (*supra* Section IV.A), examples of studies of purely legal tasks include Wistrich et al. (2015) and Kahan et al. (2016), who study the interpretation of a statute in a vignette.[29] Like KS, Kahan et al. find that judges seem to ignore legally irrelevant information (that lay people do not, *supra* Section II.A), whereas, like the ICTY experiment, Wistrich et al. find the opposite, perhaps because the former used cultural-political information that judges are sensitized to avoid assiduously (at least in the lab) while the latter used less hot-button personal traits. At the other extreme are studies using entirely abstract or at least not specifically legal tasks, which generally find judges to exhibit standard human biases (e.g., Viscusi 1999 [risk perceptions]; Guthrie et al. 2007, 2009 [CRT]; Levinson et al. 2017 [IAT]; Kneer & Bourgeois-Gironde 2017 [intentionality ascriptions]). Most other studies use vignettes that ask for a decision requiring an evaluation of the facts (as stated in the vignette) under some norm (often also stated in the vignette).

The most important line of research in this vein is the long series of vignette experiments by Chris Guthrie, Jeffrey Rachlinski, and Andrew Wistrich. We lack space to do justice to the number and richness of these experiments and refer instead to the survey by two of the authors, Rachlinski & Wistrich 2017 (also see subsequently Rachlinski & Wistrich 2018, 2021, 2022).[30] Their experiments follow a common template. They ask U.S. judges (of various levels of the federal and state judiciary) to make a decision after reading a vignette; reasons are not elicited. To trigger a

[29] Monahan & Silver's (2003) survey of the violence probability threshold that judges require to find "danger to others" randomizes the presentation of the probabilities as percentages or fractions but does so only to make their estimate independent of the presentation; with *N*=26, they lack power to study the effect of the presentation (which Rachlinski et al. 2013b find).

[30] Guthrie et al. (2001), (2007), (2009); Wistrich et al. (2005), (2015); Rachlinski et al. (2006), (2009), (2011), (2013a), (2013b), (2015). Also see similar experiments with arbitrators in Franck et al. (2017) and Helm et al. (2016).

bias, the experiment manipulates some non-legal factor, such as the gain/loss framing, hindsight, or an anchor. In some experiments, it is not clear that the manipulation is truly legally irrelevant, such that an effect of the factor can be interpreted as a bias (e.g., Kysar 2007). Nonetheless, it is fair to conclude from this body of research that judges exhibit the same heuristics and biases as other human beings, if perhaps somewhat less so (Hastie & Viscusi 1998; Viscusi 2001): judges are people, not super-humans.[31]

Of course, that does not mean that judges fall for just anything. For example, Rachlinski et al. (2006, 1254) find that bankruptcy judges' discharge decision is not influenced by an apology—which Robbennolt & Lawless (2013) confirm in a different design[32]—while Hornuf & Klöhn (2019) find that judges are not biased against someone described as a "stock market trader" compared to a "doctor" (even though lawyers predicted they would be). Wallace & Kassin (2012) find that "improper confession significantly increased … conviction rate" but also that "judges successfully overruled the confession when required to do so." The greatest surprise among the null findings is that hindsight bias appears absent in probable cause determinations (Rachlinski et al. 2011), albeit not in other settings (Anderson et al. 1993; Guthrie et al. 2001).

There is every reason to think that these heuristics and biases are universal, and manifest in the real world. Rachlinski et al. (2015) document various biases in a sample including Canadian and Dutch judges, Schweizer (2005) obtains similar results as Guthrie/Rachlinski/Wistrich with Swiss judges, Oeberst & Goeckenjan (2016) find hindsight bias in German judges' negligence assessments, and we will instantly report on anchoring in German judges. Chen et al. (2016) document the gambler's fallacy in actual asylum decisions of U.S. immigration judges, exploiting the natural experiment of random case ordering.

Some studies of criminal sentencing use longer and hence more realistic vignettes than Guthrie/Rachlinski/Wistrich. Englich & Mussweiler 2001, and Englich et al. (2005, 2006) use a four-page vignette developed with German judges to study anchoring bias in German judges. They find it every time. (So do Spamann et al. (2021) with judges from seven jurisdictions.) Some of the manipulations used (e.g., a prosecutor's demand in Englich & Mussweiler 2001) may not be entirely legally irrelevant (why else would the legal system require the prosecutor to formulate a demand?). But others evidently are, for example rolling dice (Englich et al. 2006). Skeem et al. (2020) use vignettes "formatted … as local presentence investigation reports" and Imai et al. (2023) even run a field experiment to study the impact of an algorithmic risk assessment on sentencing, which the former find to vary by the defendant's socioeconomic status and the latter find to be generally small.[33] Levinson et al. (2017) use a realistic presentencing report to study bias against Asian-Americans and Jews in sentencing by U.S. district, magistrate, and state judges, finding it in only one of the 2×3 bias-judge combinations.

There are other, less realistic sentencing studies. With Chinese judges, Yan & Lao (2022) find that male defendants receive 44 months more than female defendants for homicide but 10 (20) months *less* for fraud (drug trafficking), with no difference for robbery. Aharoni et al. (2022) find that Minnesota state judges award shorter sentences when reminded of the basic costs of incarceration such as disrupting defendants' family and the cash cost of prisons (as do Rachlinski

[31] Also see Lassiter et al. (2007) (documenting the camera-perspective effect in a mixed subject pool of judges, defense lawyers, and law enforcement officers).

[32] Robbennolt & Lawless report that "judges' assessments of debtors were influenced by apologies" but the simple difference in means that they do find is not statistically significant (see id. at 782).

[33] These experimental findings are consistent with observational studies finding that judges tend to override algorithmic recommendations in ways that dampen their impact and sometimes increases group disparities (e.g., Albright 2019; Stevenson & Doleac 2022).

et al. 2013b). Aharoni et al. correctly interpret this as bias because even though the costs may be legally relevant, the *reminder* of obvious facts should not make a difference. Similarly, reaction to a text denying the existence of free will would have been a bias—but Genschow et al. (2021) do not find any. Aspinwall et al. (2012) is a closer call because the biomechanics information they present to sentencing judges is arguably novel and thus arguably should influence judges' sentencing.

Psychologists have conducted many more experiments with judges using stripped-down vignettes. From a legal point of view, these experiments often seem unrealistic and insensitive to the subtleties of legal reasoning and fact-finding in a way that might vitiate inferences even about more general, not specifically judicial behavior. Struchiner et al. (2020) find that judges rule differently in concrete cases than they respond to abstract questions. It might be argued, however, that the abstract question encompasses other concrete cases that the judges would have decided differently than the one presented to them. Catellani et al. (2021) find that the particular counterfactual considered by an expert influences judges' decisions, which they interpret as bias. The problem is that the counterfactual might have contained legally relevant information.[34] Zenker et al. (2020) elicit different levels of agreement with an argument depending on whether it was generic or concrete (e.g., "a child" vs. "under 4 years"). While both versions of the argument were made in the context of the same case, however, they were different arguments and thus might very well deserve different responses (e.g., an argument might hold for "under 4 years" but be overbroad for a generic "child"). Rassin (2017) finds that the order in which judges see evidence changes the strength that they assign to individual pieces of evidence taken in isolation. Assessing individual pieces of evidence in isolation is not a judicial task, however, and the scale used ("not at all strong" to "absolute") is probably meaningless for both mathematical and human probabilistic reasoning. Other studies have technical issues.[35]

The only non-law experiment that is clearly not a bias experiment is Bushway et al. (2014). They test "bargaining in the shadow of the law" (Mnookin & Kornhauser 1979) in the context of plea bargaining. After reading a vignette and having had the opportunity to peruse a case file, participants—judges, prosecutors, and defense attorneys—indicated the probability of a conviction if the case went to trial, the expected sentence, and "the least severe sentence that would be acceptable for a plea deal." Bushway et al. find that their 2^4 variations of the case facts induce variation in the expected conviction and sentence, and that these in turn predict acceptable plea bargains, as predicted by the shadow bargaining model.

V. CONCLUSION

While much work has been done, the experiments reviewed here barely scratch the surface of what sustained experimental research could accomplish for the legal system and for legal theory.

[34] For example, one randomly assigned counterfactual in Catellani et al.'s medical malpractice vignette is: "If [the defendant doctor] had prescribed some simple diagnostic tests, the emergency surgery could have been avoided." The "simple diagnostic tests" were not mentioned in the rest of the vignette, and their availability probably has a bearing on whether the defendant acted negligently. A separate concern is that an experimenter effect might have been induced by their "manipulation check" (id., at 7) asking subjects to complete a sentence summarizing the experimental variation.

[35] Rassin's (2016, study 3) reminder of the possibility of false positives may not only have triggered "rational thinking," as the study intended, but may also have been interpreted as a hint that the seemingly overwhelming evidence *is* a false positive, or that the experimenter expected participants to think that way. Zenker et al. (2018, 515, 520) had very significantly different responses rates in control and treatment groups, which may have confounded results.

In particular, we agree with Engel (2022) that "[t]he extant experimental research on judicial decision-making is surprisingly little legal." Very few experiments test the effect of law (*supra* IV.A). Will the low observed levels of law-following replicate, and if so, what modifications, if any, will lead to more law following? How deep is the impact of the various heuristics and biases documented in the experimental literature? Heuristics and biases are important and ubiquitous. Nonetheless, humans have managed to fly to the moon, and at least some jurisdictions seem to run an acceptable legal system most of the time. We need to know more about how they do as well as they do, and how they could do better.

BIBLIOGRAPHY

Aharoni, Eyal, Heather M. Kleider-Offutt, Sarah F. Brosnan, and Morris B. Hoffman. "Nudges for Judges: An Experiment on the Effect of Making Sentencing Costs Explicit." *Frontiers in Psychology* 13 (2022): 1–7. https://doi.org/10.3389/fpsyg.2022.889933.

Albright, Alex. "If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions." Working paper, 2019. https://thelittledataset.com/about_files/albright_judge_score.pdf.

Anderson, John C., D. Jordan Lowe, and Philip M.J. Reckers. "Evaluation of Auditor Decisions: Hindsight Bias Effects and the Expectation Gap." *Journal of Economic Psychology* 14, no. 4 (1993): 711–37. https://doi.org/10.1016/0167-4870(93)90018-G.

Aspinwall, Lisa G., Teneille R. Brown, and James Tabery. "The Double-Edged Sword: Does Biomechanism Increase or Decrease Judges' Sentencing of Psychopaths?" *Science* 337, no. 6096 (2012): 846–49. https://doi.org/10.1126/science.1219569.

Austin, William, and Thomas A. Williams. "A Survey of Judges' Responses to Simulated Legal Cases: Research Note on Sentencing Disparity." *The Journal of Criminal Law and Criminology (1973-)* 68, no. 2 (1977): 306. https://doi.org/10.2307/1142852.

Becker, Theodore L. "A Survey Study of Hawaiian Judges: The Effect on Decisions of Judicial Role Variations." *The American Political Science Review* 60, no. 3 (1966): 677–80. https://doi.org/10.2307/1952979.

Bègue, Laurent, Oulmann Zerhouni, and Fabien Jobard. "The Role of Alcohol Intoxication on Sentencing by Judges and Laypersons: Findings From a Binational Experiment in Germany and France." *International Criminal Justice Review*, 2020, 1–13. https://doi.org/10.1177/1057567720953874.

Bornstein, Brian H., and Edie Greene. *The Jury under Fire: Myth, Controversy, and Reform*. American Psychology-Law Society Series. Oxford, UK; New York, NY: Oxford University Press, 2017.

Bourreau-Dubois, Cécile, Myriam Doriat-Duban, Bruno Jeandidier, and Jean-Claude Ray. "Do Sentencing Guidelines Result in Lower Inter-Judge Disparity ? Evidence from Framed Field Experiment (Updated Version)." Working paper, 2021. https://hal.univ-lorraine.fr/hal-03437637.

Braman, Eileen. "Cognition in the Courts." In *The Oxford Handbook of U.S. Judicial Behavior*, edited by Lee Epstein and Stefanie A. Lindquist, Online edition. 1:483–507. Oxford Academic, 2017. https://doi.org/10.1093/oxfordhb/9780199579891.013.31.

Bushway, Shawn D., Emily G. Owens, and Anne Morrison Piehl. "Sentencing Guidelines and Judicial Discretion: Quasi-Experimental Evidence from Human Calculation Errors." *Journal of Empirical Legal Studies* 9, no. 2 (2012): 291–319. https://doi.org/10.1111/j.1740-1461.2012.01254.x.

Bushway, Shawn D., Allison D. Redlich, and Robert J. Norris. "An Explicit Test of Plea Bargaining in the 'Shadow of the Trial.'" *Criminology* 52, no. 4 (2014): 723–54. https://doi.org/10.1111/1745-9125.12054.

Bystranowski, Piotr, Bartosz Janik, Maciej Próchnicki, and Paulina Skórska. "Anchoring Effect in Legal Decision-Making: A Meta-Analysis." *Law and Human Behavior* 45, no. 1 (2021): 1–23. https://doi.org/10.1037/lhb0000438.

Catellani, Patrizia, Mauro Bertolotti, Monia Vagni, and Daniela Pajardi. "How Expert Witnesses' Counterfactuals Influence Causal and Responsibility Attributions of Mock Jurors and Expert Judges." *Applied Cognitive Psychology* 35, no. 1 (2021): 3–17. https://doi.org/10.1002/acp.3720.

Chen, Benjamin Minhao, and Zhiyu Li. "The Foundations of Judicial Diffusion in China: Evidence from an Experiment." *Review of Law & Economics* 14, no. 3 (2018): 1–27. https://doi.org/10.1515/rle-2017-0008.

Chen, Daniel L., Tobias J. Moskowitz, and Kelly Shue. "Decision Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires." *The Quarterly Journal of Economics* 131, no. 3 (2016): 1181–1242. https://doi.org/10.1093/qje/qjw017.

Choi, Donghyun Danny, J. Andrew Harris, and Fiona Shen-Bayh. "Ethnic Bias in Judicial Decision Making: Evidence from Criminal Appeals in Kenya." *American Political Science Review* 116, no. 3 (2022): 1067-1080.

Chorn, Jacqueline Austin, and Margaret Bull Kovera. "Variations in Reliability and Validity Do Not Influence Judge, Attorney, and Mock Juror Decisions about Psychological Expert Evidence." *Law and Human Behavior* 43, no. 6 (2019): 542–57. https://doi.org/10.1037/lhb0000345.

City Magistrates' Courts, City of New York. *Annual Report 1916*. New York, 1917. http://hdl.handle.net/2027/njp.32101067573277.

Doob, Anthony N., and Lucien A. Beaulier. "Variation in the Exercise of Judicial Discretion with Young Offenders." *Canadian Journal of Criminology* 34, no. 1 (1992): 35–50. https://doi.org/10.3138/cjcrim.34.1.35.

Engel, Christoph. "Judicial Decision-Making: A Survey of the Experimental Evidence." Working paper, 2022. https://dx.doi.org/10.2139/ssrn.4199122.

Engel, Christoph, and Keren Weinshall. "Manna from Heaven for Judges: Judges' Reaction to a Quasi-Random Reduction in Caseload." *Journal of Empirical Legal Studies* 17, no. 4 (2020): 722–51. https://doi.org/10.1111/jels.12265.

Englich, Birte, and Thomas Mussweiler. "Sentencing Under Uncertainty: Anchoring Effects in the Courtroom." *Journal of Applied Social Psychology* 31, no. 7 (2001): 1535–51. https://doi.org/10.1111/j.1559-1816.2001.tb02687.x.

Englich, Birte, Thomas Mussweiler, and Fritz Strack. "The Last Word in Court—A Hidden Disadvantage for the Defense." *Law and Human Behavior* 29, no. 6 (2005): 705–22. https://doi.org/10.1007/s10979-005-8380-7.

———. "Playing Dice With Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making." *Personality and Social Psychology Bulletin* 32, no. 2 (2006): 188–200. https://doi.org/10.1177/0146167205282152.

Ericsson, K. Anders. "An Introduction to the Second Edition of The Cambridge Handbook of Expertise and Expert Performance: Its Development, Organization, and Content." In *The Cambridge Handbook of Expertise and Expert Performance*, edited by K. Anders Ericsson, Robert R. Hoffman, Aaron Kozbelt, and A. Mark Williams, 2nd ed., 3–20. Cambridge Handbooks in Psychology. Cambridge: Cambridge University Press, 2018. https://doi.org/10.1017/9781316480748.001.

Fischman, Joshua B. "Measuring Inconsistency, Indeterminacy, and Error in Adjudication." *American Law and Economics Review* 16, no. 1 (2014): 40–85. https://doi.org/10.1093/aler/aht011.

Franck, Susan D., Anne van Aaken, James Freda, Chris Guthrie, and Jeffrey J. Rachlinski. "Inside the Arbitrator's Mind." *Emory Law Journal* 66, no. 5 (2017): 1115–73. https://doi.org/10.31228/osf.io/ea5pm.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345, no. 6203 (2014): 1502–5. https://doi.org/10.1126/science.1255484.

Gazal-Ayal, Oren, and Raanan Sulitzeanu-Kenan. "Let My People Go: Ethnic In-Group Bias in Judicial Decisions—Evidence from a Randomized Natural Experiment." *Journal of Empirical Legal Studies* 7, no. 3 (2010): 403–28. https://doi.org/10.1111/j.1740-1461.2010.01183.x.

Genschow, Oliver, Heinz Hawickhorst, Davide Rigoni, Ellen Aschermann, and Marcel Brass. "Professional Judges' Disbelief in Free Will Does Not Decrease Punishment." *Social Psychological and Personality Science* 12, no. 3 (2021): 357–62. https://doi.org/10.1177/1948550620915055.

Green, Donald P., and Dane R. Thorley. "Field Experimentation and the Study of Law and Policy." *Annual Review of Law and Social Science* 10, no. 1 (2014): 53–72. https://doi.org/10.1146/annurev-lawsocsci-110413-030936.

Guthrie, Chris, Jeffrey J. Rachlinski, and Andrew J. Wistrich. "Inside the Judicial Mind." *Cornell Law Review* 86, no. 4 (2001): 777–830.

———. "Blinking on the Bench: How Judges Decide Cases." *Cornell Law Review* 93, no. 1 (2007): 1–43.

———. "The 'Hidden Judiciary': An Empirical Examination of Executive Branch Justice." *Duke Law Journal* 58, no. 7 (2009): 1477–1530.

Harris, Allison P., and Maya Sen. "Bias and Judging." *Annual Review of Political Science* 22 (2019): 241–59. https://doi.org/10.1146/annurev-polisci-051617-090650.

Hastie, Reid, and W. Kip Viscusi. "What Juries Can't Do Well: The Jury's Performance As a Risk Manager." *Arizona Law Review* 40, no. 3 (1998): 901–22.

Helm, Rebecca K., Andrew J. Wistrich, and Jeffrey J. Rachlinski. "Are Arbitrators Human?" *Journal of Empirical Legal Studies* 13, no. 4 (2016): 666–92. https://doi.org/10.1111/jels.12129.

Hofer, Paul J. "*United States v. Booker* as a Natural Experiment: Using Empirical Research to Inform the Federal Sentencing Policy Debate." *Criminology & Public Policy* 6, no. 3 (2007): 433–60. https://doi.org/10.1111/j.1745-9133.2007.00446.x.

Hood, Roger. *Sentencing the Motoring Offender. A Study of Magistrates' Views and Practices*. London: Heinemann Educational Books, 1972.

Hornuf, Lars, and Lars Klöhn. "Do Judges Hate Speculators?" *European Journal of Law and Economics* 47, no. 2 (2019): 147–69. https://doi.org/10.1007/s10657-018-09608-z.

Hubbard, William H. J. "The Effects of Twombly and Iqbal." *Journal of Empirical Legal Studies* 14, no. 3 (2017): 474–526. https://doi.org/10.1111/jels.12153.

Imai, Kosuke, Zhichao Jiang, D James Greiner, Ryan Halen, and Sooahn Shin. "Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment." *Journal of the Royal Statistical Society Series A: Statistics in Society*, 2023. https://doi.org/10.1093/jrsssa/qnad010.

Kahan, Dan M., David Hoffman, Danieli Evans, Eugene Lucci, and Katherine Cheng. "'Ideology' or 'Situation Sense'? An Experimental Investigation of Motivated Reasoning and Professional Judgment." *University of Pennsylvania Law Review* 164, no. 2 (2016): 349–439.

Kahneman, Daniel, and Gary Klein. "Conditions for Intuitive Expertise A Failure to Disagree." *American Psychologist* 64 (2009): 515–26. https://doi.org/10.1037/a0016755.
Kalven, Harry, and Hans Zeisel. *The American Jury*. Boston: Little, Brown, 1966.
Kennedy, Duncan. *A Critique of Adjudication (Fin de Siècle)*. Cambridge, MA: Harvard University Press, 1998.
Kihlstrom, John F. "Ecological Validity and 'Ecological Validity'." *Psychological Science* 16 (2021): 466-71.
Klein, David. "Law in Judicial Decision-Making." In *The Oxford Handbook of U.S. Judicial Behavior*, edited by Lee Epstein and Stefanie A. Lindquist, Online edition. 1:236–52. Oxford Academic, 2017. https://doi.org/10.1093/oxfordhb/9780199579891.013.5.
Klerman, Daniel, and Holger Spamann. "Law Matters – Less Than We Thought." *The Journal of Law, Economics, and Organization*, 2024. https://doi.org/10.1093/jleo/ewac008.
Kneer, Markus, and Sacha Bourgeois-Gironde. "Mens Rea Ascription, Expertise and Outcome Effects: Professional Judges Surveyed." *Cognition* 169 (2017): 139–46. https://doi.org/10.1016/j.cognition.2017.08.008.
Kopas, Jacob, and Dane Thorley. "Experiments in the Court: The Legal and Ethical Challenges of Running Randomized Field Experiments in the Courtroom." Working paper, 2018. https://doi.org/10.2139/ssrn.2994298.
Krasno, Jonathan S., Donald P. Green, Costas Panagopoulos, Dane Thorley, and Michael Schwam-Baird. "Campaign Donations, Judicial Recusal, and Disclosure: A Field Experiment." *The Journal of Politics* 83, no. 4 (2021): 1844–50. https://doi.org/10.1086/715069.
Kysar, Douglas A. "The Jurisprudence of Experimental Law and Economics." *Journal of Institutional and Theoretical Economics* 163, no. 1 (2007): 187–98. https://doi.org/10.1628/093245607780182017.
Landsman, Stephan, and Richard F. Rakos. "A Preliminary Inquiry into the Effect of Potentially Biasing Information on Judges and Jurors in Civil Litigation." *Behavioral Sciences & the Law* 12, no. 2 (1994): 113–26. https://doi.org/10.1002/bsl.2370120203.
Lassiter, G. Daniel, Shari Seidman Diamond, Heather C. Schmidt, and Jennifer K. Elek. "Evaluating videotaped confessions: Expertise provides no defense against the camera-perspective effect." *Psychological Science*, 18, no. 3 (2007): 224–26. https://doi.org/10.1111/j.1467-9280.2007.01879.x.
Lemon, Nigel. "Training, Personality and Attitudes as Determinants of Magistrates' Sentencing." *British Journal of Criminology* 14, no. 1 (1974): 34–48. https://doi.org/10.1093/oxfordjournals.bjc.a046509.
Levinson, Justin D., Mark W. Bennett, and Koichi Hioki. "Judging Implicit Bias: A National Empirical Study of Judicial Stereotypes." *Florida Law Review* 69, no. 1 (2017): 63–113.
Lindholm, Torun. "Who Can Judge the Accuracy of Eyewitness Statements? A Comparison of Professionals and Lay-Persons." *Applied Cognitive Psychology* 22, no. 9 (2008): 1301–14. https://doi.org/10.1002/acp.1439.
Liu, John Zhuang, Lars Klöhn, and Holger Spamann. "Precedents and Chinese Judges: An Experiment." *The American Journal of Comparative Law* 69, no. 1 (2021): 93–135. https://doi.org/10.1093/ajcl/avab009.
Liu, John Zhuang, and Xueyao Li. "Legal Techniques for Rationalizing Biased Judicial Decisions: Evidence from Experiments with Real Judges." *Journal of Empirical Legal Studies* 16, no. 3 (2019): 630–70. https://doi.org/10.1111/jels.12229.

Liu, Zhuang. "Does Reason Writing Reduce Decision Bias? Experimental Evidence from Judges in China." *The Journal of Legal Studies* 47, no. 1 (2018): 83–118. https://doi.org/10.1086/696879.

Llewellyn, Karl N. "A Realistic Jurisprudence -- The Next Step." *Columbia Law Review* 30, no. 4 (1930): 431–65. https://doi.org/10.2307/1114548.

Lynch, H. Fernandez, D. J. Greiner, and I. G. Cohen. "Overcoming Obstacles to Experiments in Legal Practice." *Science* 367, no. 6482 (2020): 1078–80. https://doi.org/10.1126/science.aay3005.

Macdonald, Scott, Patricia Erickson, and Barbara Allen. "Judicial Attitudes in Assault Cases Involving Alcohol or Other Drugs." *Journal of Criminal Justice* 27, no. 3 (1999): 275–86. https://doi.org/10.1016/S0047-2352(98)00065-8.

Macleod, James A. "Ordinary Causation: A Study in Experimental Statutory Interpretation." *Indiana Law Journal* 94, no. 3 (2019): 957–1029.

———. "Finding Original Public Meaning." *Georgia Law Review* 56, no. 1 (2021): 1–79.

McQuiston-Surrett, Dawn, and Michael J. Saks. "The Testimony of Forensic Identification Science: What Expert Witnesses Say and What Factfinders Hear." *Law and Human Behavior* 33, no. 5 (2009): 436–53. https://doi.org/10.1007/s10979-008-9169-1.

Miller, Andrea L. "Expertise Fails to Attenuate Gendered Biases in Judicial Decision-Making." *Social Psychological and Personality Science* 10, no. 2 (2019): 227–34. https://doi.org/10.1177/1948550617741181.

Mnookin, Robert H., and Lewis Kornhauser. "Bargaining in the Shadow of the Law: The Case of Divorce." *The Yale Law Journal* 88, no. 5 (1979): 950–97. https://doi.org/10.2307/795824.

Monahan, John, and Eric Silver. "Judicial Decision Thresholds for Violence Risk Management." *International Journal of Forensic Mental Health* 2, no. 1 (2003): 1–6. https://doi.org/10.1080/14999013.2003.10471174.

Nelson, Leif D., Joseph Simmons, and Uri Simonsohn. "Psychology's Renaissance." *Annual Review of Psychology* 69, no. 1 (2018): 511–34. https://doi.org/10.1146/annurev-psych-122216-011836.

Oeberst, Aileen, and Ingke Goeckenjan. "When Being Wise After the Event Results in Injustice: Evidence for Hindsight Bias in Judges' Negligence Assessments." *Psychology, Public Policy, and Law* 22, no. 3 (2016): 271–79.

Palys, T. S., and Stan Divorski. "Explaining Sentence Disparity." *Canadian Journal of Criminology* 28, no. 4 (1986): 347–62. https://doi.org/10.3138/cjcrim.28.4.347.

Pantazi, Myrto, Olivier Klein, and Mikhail Kissine. "Is Justice Blind or Myopic? An Examination of the Effects of Meta-Cognitive Myopia and Truth Bias on Mock Jurors and Judges." *Judgment and Decision Making* 15, no. 2 (2020): 214–29. https://doi.org/10.1017/S1930297500007361.

Rachlinski, Jeffrey J., Chris Guthrie, and Andrew J. Wistrich. "Inside the Bankruptcy Judge's Mind." *Boston University Law Review* 86, no. 5 (2006): 1227–66.

———. "Probable Cause, Probability, and Hindsight: Probable Cause, Probability, and Hindsight." *Journal of Empirical Legal Studies* 8, no. S1 (2011): 72–98. https://doi.org/10.1111/j.1740-1461.2011.01230.x.

———. "Contrition in the Courtroom: Do Apologies Affect Adjudication?" *Cornell Law Review* 98, no. 5 (2013a): 1189–1244.

Rachlinski, Jeffrey J., Sheri Lynn Johnson, Andrew J. Wistrich, and Chris Guthrie. "Does Unconscious Racial Bias Affect Trial Judges?" *Note Dame Law Review* 84, no. 3 (2009): 1195–1246.

Rachlinski, Jeffrey J., and Andrew J. Wistrich. "Judging the Judiciary by the Numbers: Empirical Research on Judges." *Annual Review of Law and Social Science* 13, no. 1 (2017): 203–29. https://doi.org/10.1146/annurev-lawsocsci-110615-085032.

———. “Gains, Losses, and Judges: Framing and the Judiciary.” *Notre Dame Law Review* 94, no. 2 (2018): 521–82.
———. “Benevolent Sexism in Judges.” *San Diego Law Review* 58, no. 1 (2021): 101–41.
———. “Judging Autonomous Vehicles.” *Yale Journal of Law and Technology* (forthcoming), 2022. https://doi.org/10.2139/ssrn.3806580.
Rachlinski, Jeffrey J., Andrew J. Wistrich, and Chris Guthrie. “Altering Attention in Adjudication.” *UCLA Law Review* 60, no. 6 (2013b): 1586–1618.
———. “Can Judges Make Reliable Numeric Judgments? Distorted Damages and Skewed Sentences.” *Indiana Law Journal* 90, no. 2 (2015): 695–739.
Ramji-Nogales, Jaya, Andrew I. Schoenholtz, and Philip G. Schrag. “Refugee Roulette: Disparities in Asylum Adjudication.” *Stanford Law Review* 60, no. 2 (2007): 295–412.
Rassin, Eric. “Rational Thinking Promotes Suspect-Friendly Legal Decision Making.” *Applied Cognitive Psychology* 30, no. 3 (2016): 460–64. https://doi.org/10.1002/acp.3198.
———. “Initial Evidence for the Assimilation Hypothesis.” *Psychology, Crime & Law* 23, no. 10 (2017): 1010–20. https://doi.org/10.1080/1068316X.2017.1371307.
Redding, Richard E., and N. Dickon Reppucci. “Effects of Lawyers’ Socio-Political Attitudes on Their Judgments of Social Science in Legal Decision Making.” *Law and Human Behavior* 23, no. 1 (1999): 31–54. https://doi.org/10.1023/A:1022322706533.
Robbennolt, Jennifer K. “Punitive Damage Decision Making: The Decisions of Citizens and Trial Court Judges.” *Law and Human Behavior* 26, no. 3 (2002): 315–41. https://doi.org/10.1023/A:1015376421813.
———. “Apologies and Legal Settlement: An Empirical Examination.” *Michigan Law Review* 102, no. 3 (2003): 460–516. https://doi.org/10.2307/3595367.
———. “Evaluating Juries by Comparison to Judges: A Benchmark for Judging?” *Florida State University Law Review* 32, no. 2 (2005): 469–509.
Robbennolt, Jennifer K., and Robert M. Lawless. “Bankrupt Apologies.” *Journal of Empirical Legal Studies* 10, no. 4 (2013): 771–96. https://doi.org/10.1111/jels.12027.
Rosenblatt, Abram, Jeff Greenberg, Sheldon Solomon, Tom Pyszczynski, and Deborah Lyon. “The Effects of Mortality Salience on Reactions to Those Who Violate or Uphold Cultural Values.” *Journal of Personality and Social Psychology* 57, no. 4 (1989): 681-690.
Schauer, Frederick. “Is There a Psychology of Judging?” In *The Psychology of Judicial Decision Making*, edited by David E. Klein and Gregory Mitchell, Online edition. 103–20. Oxford Academic, 2010. https://doi.org/10.1093/acprof:oso/9780195367584.003.0007.
Schmittat, Susanne M., and Birte Englich. “If You Judge, Investigate! Responsibility Reduces Confirmatory Information Processing in Legal Experts.” *Psychology, Public Policy, and Law* 22, no. 4 (2016): 386–400. https://doi.org/10.1037/law0000097.
Schweizer, Mark Daniel. “Kognitive Täuschungen Vor Gericht - Eine Empirische Studie,” PhD diss., University of Zurich, 2005. https://doi.org/10.5167/uzh-165152.
Simon, Dan. “On Juror Decision Making: An Empathic Inquiry.” *Annual Review of Law and Social Science* 15, no. 1 (2019): 415–35. https://doi.org/10.1146/annurev-lawsocsci-101518-042658.
Skeem, Jennifer, Nicholas Scurich, and John T. Monahan. “Impact of Risk Assessment on Judges’ Fairness in Sentencing Relatively Poor Defendants.” *Law and Human Behavior* 44, no. 1 (2020): 51–59.
Sonnemans, Joep, and Frans van Dijk. “Errors in Judicial Decisions: Experimental Results.” *Journal of Law, Economics, and Organization* 28, no. 4 (2012): 687–716. https://doi.org/10.1093/jleo/ewq019.

Spamann, Holger. "Comment on 'Temperature and Decisions: Evidence from 207,000 Court Cases.'" *American Economic Journal: Applied Economics* 14, no. 4 (2022): 519–28. https://doi.org/10.1257/app.20200118.

Spamann, Holger, and Lars Klöhn. "Justice Is Less Blind, and Less Legalistic, Than We Thought: Evidence from an Experiment with Real Judges." *Journal of Legal Studies* 45, no. 2 (2016): 255–80. https://doi.org/10.1086/688861.

———. "Can Law Students Replace Judges in Experiments of Judicial Decision-Making?" Working paper, 2023. https://papers.ssrn.com/abstract=4362199.

Spamann, Holger, Lars Klöhn, Christophe Jamin, Vikramaditya Khanna, John Zhuang Liu, Pavan Mamidi, Alexander Morell, and Ivan Reidel. "Judges in the Lab: No Precedent Effects, No Common/Civil Law Differences." *Journal of Legal Analysis* 13, no. 1 (2021): 110–26. https://doi.org/10.1093/jla/laaa008.

Spellmann, Barbara A. "Judges, Expertise, and Analogy." In *The Psychology of Judicial Decision Making*, edited by David E. Klein and Gregory Mitchell, Online edition., 149–64. Oxford Academic, 2010. https://doi.org/10.1093/acprof:oso/9780195367584.003.0010.

Stevenson, Megan T., and Jennifer L. Doleac. "Algorithmic Risk Assessment in the Hands of Humans." Working paper, 2022. https://ssrn.com/abstract=3489440.

Struchiner, Noel, Guilherme da F. C. F. de Almeida, and Ivar R. Hannikainen. "Legal Decision-Making and the Abstract/Concrete Paradox." *Cognition* 205 (2020): 1–15. https://doi.org/10.1016/j.cognition.2020.104421.

Teichman, Doron, and Eyal Zamir. "Judicial Decision-Making: A Behavioral Perspective." In *The Oxford Handbook of Behavioral Economics and the Law*, edited by Eyal Zamir and Doron Teichman, Online edition., 663–702. Oxford Academic, 2014. https://doi.org/10.1093/oxfordhb/9780199945474.013.0026.

Thompson, Neil, Brian Flanagan, Edana Richardson, Brian McKenzie, and Xueyun Luo. "Trial by Internet: A Randomized Field Experiment on Wikipedia's Influence on Judges' Legal Reasoning." *Cambridge Handbook of Experimental Jurisprudence* (forthcoming), 2022. https://dx.doi.org/10.2139/ssrn.4174200.

Tobia, Kevin. "Legal Concepts and Legal Expertise." Working paper, 2020. https://doi.org/10.2139/ssrn.3536564.

Tobia, Kevin, Brian G. Slocum, and Victoria Nourse. "Statutory Interpretation from the Outside." *Columbia Law Review* 122, no. 1 (2022): 213–329.

Vidmar, Neil, and Valerie P. Hans. *American Juries: The Verdict*. 1st American hardcover ed. Amherst, N.Y.: Prometheus Books, 2007.

Viscusi, W. Kip. "How Do Judges Think about Risk?" *American Law and Economics Review* 1, no. 1 (1999): 26–62. https://doi.org/10.1093/aler/1.1.26.

———. "Jurors, Judges, and the Mistreatment of Risk by the Courts." *The Journal of Legal Studies* 30, no. 1 (2001): 107–42. https://doi.org/10.1086/468113.

Wallace, D. Brian, and Saul M. Kassin. "Harmless Error Analysis: How Do Judges Respond to Confession Errors?" *Law and Human Behavior* 36, no. 2 (2012): 151–57. https://doi.org/10.1037/h0093975.

Wessel, Ellen, Guri C. B. Drevland, Dag Erik Eilertsen, and Svein Magnussen. "Credibility of the Emotional Witness: A Study of Ratings by Court Judges." *Law and Human Behavior* 30, no. 2 (2006): 221–30. https://doi.org/10.1007/s10979-006-9024-1.

Wissler, Roselle L., Allen J. Hart, and Michael J. Saks. "Decisionmaking About General Damages: A Comparison of Jurors, Judges, and Lawyers." *Michigan Law Review* 98, no. 3 (1999): 751–826.

Wistrich, Andrew J., Chris Guthrie, and Jeffrey J. Rachlinski. "Can Judges Ignore Inadmissible Information? The Difficulty of Deliberately Disregarding." *University of Pennsylvania Law Review* 153, no. 4 (2005): 1251–1345. https://doi.org/10.2307/4150614.

Wistrich, Andrew J., Jeffrey J. Rachlinski, and Chris Guthrie. "Heart Versus Head: Do Judges Follow the Law or Follow Their Feelings?" *Texas Law Review* 93 (2015): 855–923.

Yan, Shi, and Jiaqi Lao. "Sex Disparities in Sentencing and Judges' Beliefs: A Vignette Approach." *Victims & Offenders* 17, no. 4 (2022): 597–619. https://doi.org/10.1080/15564886.2021.1947427.

Yang, Crystal S. "Free at Last? Judicial Discretion and Racial Disparities in Federal Sentencing." *The Journal of Legal Studies* 44, no. 1 (2015): 75–111. https://doi.org/10.1086/680989.

Zeiler, Kathryn. "Cautions on the Use of Economics Experiments in Law." *Journal of Institutional and Theoretical Economics* 166, no. 1 (2010): 178–93. https://doi.org/10.1628/093245610790711483.

Zenker, Frank, Christian Dahlman, Rasmus Bååth, and Farhan Sarwar. "Reasons Pro et Contra as a Debiasing Technique in Legal Contexts." *Psychological Reports* 121, no. 3 (2018): 511–26. https://doi.org/10.1177/0033294117729807.

Zenker, Frank, Christian Dahlman, Sverker Sikström, Lena Wahlberg, and Farhan Sarwar. "Generalization in Legal Argumentation." *Journal of Forensic Psychology Research and Practice* 20, no. 1 (2020): 80–99. https://doi.org/10.1080/24732850.2019.1689782.