JUDGES IN THE LAB: NO PRECEDENT EFFECTS, NO COMMON/CIVIL LAW DIFFERENCES

Holger Spamann*, Lars Klöhn††, Christophe Jamin‡, Vikramaditya Khanna§, John Zhuang Liu*, Pavan Mamidi‖, Alexander Morell** and Ivan Reidel††

* Lawrence R. Grove Professor, Harvard Law School, Cambridge, MA. Email: hspamann@law.harvard.edu. Spamann is a co-editor of the JLA. This article was selected for publication by the Editors-in-Chief, after undergoing peer review. Spamann was not involved in the selection process.
† Professor, Humboldt Universität zu Berlin, Faculty of Law, Berlin, Germany.
‡ Professor, Sciences Po Law School, Paris, France.
§ William W. Cook Professor, University of Michigan Law School, Ann Arbor, MI.
‖ Assistant Professor, The Chinese University of Hong Kong Shenzhen, School of Management and Economics, Shenzhen, PRC.
** Professor, Universität Mannheim, Department of Law, Mannheim, Germany.
†† CEO, Sinfonik Inc., New York, NY.

We thank FLAME University, Harvard Law School, Harvard Law School’s Summer Research program, Humboldt Universität zu Berlin, Ludwig-Maximilians-Universität, the Max Planck Institute for Collective Goods, Sciences Po, and Universidad Torcuato di Tella for financial support; the Max Planck Institute for Collective Goods for use of their mobile lab; Roland Ramthun of Docustorm for programming the experimental interface; Brendan Halpin for help with the Stata package SADI (and for programming it in the first place); Alex Whiting for suggesting the aiding and abetting controversy; Damien Charlotin, Leandro Dias, and Máximo Langer for information about national laws; Gabriela Haymes, Henrique Duchini, Allana Pereira, Andrioli Soares, Chen Gu, Linda Hao (Langscale Translation), Aurelien Bernard, Damien Charlotin, Jakob Joehmann, and Lola Witt for translating the experimental materials into local languages; Marcelo Moreno Bonassa, Etiene Coelho Martins, Linda Hao (Langscale Translation), Damien Charlotin, Mareike Grundmann, and Andrew Hammel for translating the judgment reasons from local languages into English; James Foy, Andres Constantin, Victoria Arredondo Cívico, Amanda Bernsteine, Nicole Varela, Julian Franke, Oliver Schöen, Martin Sternberg, Philipp Schmidt, Andrés Parrado, Shardul Vaidya, Howard Duan, Molly Eskridge, Tyler Good-Cohn, and JohneHenre Rzodkiewicz for excellent research assistance and administrative support; Judge Marta Capalbo, Judge Duilio Campora, and Mariano Sigman (Argentina), Judge Bruno Bodart, Estevao Gomes, Judge Isabela Isa, and HLSBSA (Brazil), Xueyao Li (China), Justice Dominique Lottin, Judge Pénélope Postel-Vinay, and the ENM (France), Deutsche Richterakademie (Germany), Ranita Nagar, Kalpesh Kumar, Gujarat National Law University, and the Chief Justice of the Gujarat High Court (India), Judge Nancy Gertner, John Manning, Denise Neary, and the FJC (USA), for accommodating and/or supporting the respective sessions at which we collected our data; and above all the judges for their participation. For comments, we thank participants and commentators at the American Society of Comparative Law’s Special Committee on Comparative Law and the Social Sciences, the Max Planck

© The Author(s) 2021. Published by Oxford University Press on behalf of The John M. Olin Center for Law, Economics and Business at Harvard Law School.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
doi:10.1093/jla/laa008
ABSTRACT

In our lab, 299 real judges from seven major jurisdictions (Argentina, Brazil, China, France, Germany, India, and USA) spend up to fifty-five minutes to judge an international criminal appeals case and determine the appropriate prison sentence. The lab computer (i) logs their use of the documents (briefs, statement of facts, trial judgment, statute, precedent) and (ii) randomly assigns each judge (a) a horizontal precedent disfavoring, favoring, or strongly favoring defendant, (b) a sympathetic or an unsympathetic defendant, and (c) a short, medium, or long sentence anchor. Document use and written reasons differ between countries but not between common and civil law. Precedent effect is barely detectable and estimated to be less, and bounded to be not much greater than, that of legally irrelevant defendant attributes and sentence anchors.

INTRODUCTION

We use a novel type of evidence and a novel experimental manipulation in a novel setting to investigate two central questions in comparative law and judicial decision-making, respectively. First, is it true that common and civil lawyers in general and judges in particular think differently? (Common lawyers are those from England and her former colonies, civil lawyers the rest of the world.). Prominent writers assert such differences (Zweigert & Kötz 1998 §5.III.2; Legrand 1996, 1999: “irreducible epistemological chasm”) and link them to intercountry differences in legal rules and economic outcomes (La Porta, Lopez-de-Silanes, & Shleifer 2008; cf. Spamann 2015), but the assertion has never been tested rigorously. We track judges’ document use at high resolution and find no support for the assertion. Second, are judges’ decisions causally affected by primary legal sources, specifically by horizontal precedent (i.e. precedent set by a different panel of the same court)? Using random assignment of precedent, we find no causal effect, at least none much larger than two biases studied in the prior literature, which we replicate. We do all of this in a
laboratory setting that resembles real-world judicial decision-making, at least much more so than previous studies: our participants are real judges from seven countries who spend almost an hour to decide the same fully briefed legal case.

Judicial opinions—the official reasons judges write for their decisions—would suggest the opposite answers to both questions. The style of judicial opinions differs systematically between countries, ranging from the narrowly technical, terse, and syllogistic style of the French Cour de Cassation to the more liberal, discursive, and expansive style of the U.S. Supreme Court (e.g. Lasser 2004), and many have seen systematic differences between common and civil law (e.g. Wetter 1960; Kötz 1982). Similarly, most courts extensively cite precedent, including horizontal precedent, even in jurisdictions that do not have an explicit norm of binding precedent like the common law (stare decisis) (Summers & Taruffo 1991 §VI). As evidence of judicial thinking, however, judicial opinions are at least problematic. To maintain their legitimacy, judges must project a certain image, which may differ by country, but what really drives judges’ decisions is a different matter (e.g. Gény 1899; Kantorowicz 1906; Frank 1930; Llewellyn 1930; Kennedy 1998; Simon 1998; Lasser 2004; Epstein, Landes, & Posner 2013).

To answer the first question—differences between common and civil law thinking—without relying on judicial opinions, we generate a novel type of evidence: judges’ document use while they work on deciding a case. This “document path” process imaging technique reveals judicial thought structures in a similar way as magnetic resonance imagings and other brain imaging techniques reveal brain structures: the image is partial and requires statistical inference, but it is nevertheless highly informative. We analyze similarities between document paths using sequence analysis (edit distance), a method that we show to be very powerful in detecting even small differences between groups of judges. Nevertheless, we detect only differences between individual countries but not between common and civil law judges. Common and civil lawyers do not think differently after all, at least those in our sample at the resolution of our document path image.

To answer the second question—the effect of horizontal precedent—we randomly assign each judge one of three horizontal precedents. While this experimental manipulation may seem basic, prior experimental work with judges

---

1 At a minimum, judges have more discretion than their opinions admit: A lower bound on legal indeterminacy and judicial error is set by inconsistency between judges deciding the same cases on a panel (e.g. justices on the U.S. Supreme Court) or deciding random draws from the same distribution of cases (e.g. judges on U.S. immigration courts) (Fischman, 2014). How judges use this discretion correlates with their background and ideology (e.g. Rachlinski & Wistrich 2017; Harris & Sen 2019).
(e.g. Guthrie, Rachlinski, & Wistrich 2001, 2007; Englich, Mussweiler, & Strack 2006; Wistrich et al. 2014; Kahan et al. 2016) has only studied the causal effect of non-law, i.e. biases.2 One reason explaining the prior focus on biases may be that manipulating the law is difficult with expert judges, as we discuss below. Another reason may be that the effect of the law seems beyond doubt in many situations (see ‘Discussion’ Section). However, there are examples of blatant judicial disregard of the law—even of statutes (Henderson & Hubbard 2015)—and in many situations, the causal—as opposed to rhetorical—effect of law is very much an open question. Moreover, estimating the size of the law effect, if any, will put into perspective the size of biases documented in the literature and vice versa. We replicate the anchoring and sympathy biases observed in prior studies and show that their estimated effect is larger than the estimated effect of precedent (with a 95% confidence interval ruling out significantly larger precedent effects).

We create a unique setting to study both questions: 299 real judges from 7 key jurisdictions spend up to 55 minutes deciding—with written reasons—a fully briefed appeals case. From the perspective of comparative law, this setting is unique simply because comparative law has never used laboratory or experimental methods—we are not aware of any other study that has observed lawyers from multiple jurisdictions under controlled conditions. We recruited our participants at comparable professional education seminars in jurisdictions representing half the world population and playing key roles in the common/civil law taxonomy. France (n = 43) and Germany (n = 74) are the historical and, by most accounts, current center of the civil law world (e.g. Zweigert & Kötz 1998). Others see the current center of the civil law in Latin America (Merryman & Pérez-Perdomo 2007), of which Brazil (n = 33) is the largest and Argentina (n = 31) the historically dominant jurisdiction. The USA (n = 29) is one of the two current centers of the common law world (our attempts to recruit participants in the other, historical center of the common law, England, were unsuccessful). China (n = 47) and India (n = 42) are the world’s largest countries and belong to the wider civil and common law, respectively (on China, see (Liu, Klöhn, & Spamann 2020). Informed consent was obtained. See Supplementary Information S1 for more information on venues and recruitment.

From the perspective of experimental studies of judicial decision-making, our setting offers an unusually high degree of realism. Most other studies do not have judges as research subjects, and those that do provide only a vignette

2 Spamann & Klohn (2016) and Liu et al. (2020) report partial results of the present study for the USA and China, respectively. A parallel study by one of us, Klerman & Spamann (2019), is discussed below.
summary of the decision context and do not ask for reasons. These other studies may fail to capture key, outcome-determinative features of judicial decision-making in the real world. Judges are highly trained and selected professionals. They operate in an environment designed to elucidate all sides of the case through adversarial argument, emphasize the importance of objectivity through decorum, provide time for reflection, and create accountability, principally by the obligation to provide reasons for the decision (Llewellyn 1930; Spellman & Schauer 2012; Kahan 2015; Spamann & Klöhn 2016). Our design includes all of these features.

Our study casts participants in the role of a judge on the Appeals Chamber of the International Criminal Tribunal for the Former Yugoslavia (ICTY) (see Supplementary Information S2 for more details). Participants’ task is to decide a fictionalized and streamlined version of a real ICTY case, Prosecutor v. Perišić. The legal question in the case is the meaning of aiding and abetting war crimes, which is criminalized but not defined in Article 7(1) of the ICTY Statute. Participants have fifty-five minutes to decide, with brief reasons, whether to reverse or affirm the defendant’s conviction by the ICTY’s lower chamber (“trial judgment”). Subsequently, participants are asked to set an appropriate sentence, i.e. prison term (if they decided to reverse the conviction, they are told to imagine they had been outvoted on a panel). The computer records the document passages live on a participant’s screen at any given time. The documents provided (in participants’ local language) are briefs for both parties and a statement of agreed facts (both fictional) as well as the ICTY Statute, the trial judgment, and one prior decision of the Appellate Chamber, i.e. horizontal precedent (all real). See Supplementary Information S2 for more information on the task and materials (which are reproduced in full in the Supplementary Appendix) and Supplementary Information S4 for more details on the data collected.

Unbeknownst to participants, each participant is randomly assigned one of two defendants, one of three precedents, and—in the sentencing part—one of three anchors in a $2 \times 3 \times 3$ factorial experiment (see Supplementary Information S3 for more details). The defendant is either sympathetic—a regretful, conciliatory Croat—or unsympathetic—a hateful, nationalist Serb; these attributes are strictly legally irrelevant for purposes of determining whether a crime has been committed even though they may be taken into account for sentencing, i.e. for setting the appropriate penalty (e.g. Nadler &

---

3 The closest other studies we are aware of in this respect are Englich & Mussweiler (2001) and Englich et al. (2006), who provide a four-page description of a rape case to German judges spending about fifteen minutes on a sentencing task; however, in Germany, a judge would never sentence a defendant without having first determined guilt during a full trial. One of us uses a similarly realistic design in a subsequent study discussed below (Klerman & Spamann 2019).
McDonnell 2012). The precedent either supports affirmance (Affirm) or supports reversal weakly (reverse) or strongly (REVERSE). REVERSE is a disguised version of the actual decision in Prosecutor v. Perišić, i.e. it concerns the exact same facts and is thus the strongest precedent imaginable. 4 Affirm is the explicit, lengthy refutation of REVERSE by another panel of the Appeals Chamber and hence very strong precedent in the opposite direction even while dealing with somewhat different facts (we redacted Affirm to hide the existence of REVERSE; in our redaction, Affirm addresses a position advanced by the defense). reverse defines aiding and abetting in terms that would exclude our defendant’s behavior; however, this definition is neither discussed at length nor outcome-determinative in reverse, limiting its precedential weight. Finally, the sentencing task is explained using an example of ten, twenty-five, or forty-years (anchor). Prior research on judicial biases has shown judges to be affected by anchors, specifically in sentencing (e.g. Englich and Mussweiler 2001; Guthrie et al. 2001; Englich, Mussweiler, & Strack 2006; Rachlinski, Wistrich, & Guthrie 2015), and by party sympathies (Rachlinski et al. 2009; Wistrich et al. 2014). See Supplementary Information S3 for more information on the experimental treatments.

As in our study, many domestic cases are decided by a single judge without a hearing on a limited record and in about one hour (Spamann & Klöhn 2016 §5.2). Our reason to use an ICTY case is that it is neutral between participants’ national backgrounds, while its criminal law subject matter is familiar to participants from their national law. The ecological validity of our study hinges on whether participants understand the case and approach it with their judicial mindset. Participants’ written reasons and behavior in the lab suggest they do. With very few exceptions (the exclusion of which does not affect results), participants’ reasons are coherent and mostly adopt a specifically judicial diction (see Supplementary Information S1). In the lab, participants approached the task with the utmost concentration and seriousness. Four actual or potential participants even expressed reservations about “deciding” a case of this gravity in such a short time. Most participants finished early, but some had to be asked repeatedly to finish writing their reasons. Subsequent work by one of us (Klerman & Spamann 2019) replicates the absence of a (strong) law effect in a similarly realistic design with a purely domestic lower court case, eliminating concerns of lacking realism of an international appeals case.

4 We did not use REVERSE in the first country where we ran the study, the USA, because we had expected that reverse would be strong enough to generate a precedent effect. We then added REVERSE in the other six countries.
RESULTS

Comparative Judicial Thinking

Figure 1 plots our main comparative evidence, the judges’ document view paths by country. Each judge’s time scale is normalized by that judge’s total time, so that all paths are of the same length. This normalization suppresses variation in length between individual judges (mean 35 minutes, SD 10 minutes) and between countries (Argentina 37, Brazil 32, China 28, France 37, Germany 36, India 40, USA 36) that we believe to be orthogonal to our question of interest, which is how judges think, i.e. how they form their view of a case, not how long it takes them to form that view.

At a high level and relative to the range of possible paths, the distribution of judges’ paths looks similar in all countries. Judges everywhere tend to begin with the facts and the briefs (which the study instructions recommended reading in full) before examining the trial judgment, with brief glances at the statute throughout and longer examinations of the precedent later. This is only a tendency, though: individual judges do not all start and end with the same documents nor move through the documents at the same rhythm; there is even more heterogeneity at the paragraph level, which we have found too variable for insight and thus do not report. To the naked eye, the heterogeneity is mostly within country; there are no obvious country patterns.

For a more nuanced, formal analysis, the first question we need to address is the choice of analysis method. Since our data are of a novel type, there is no established method for analyzing them in comparative law and related fields. Nor is the widespread belief that common and civil lawyers think differently articulated in specific testable hypotheses. Simple multivariate ANOVA with the frequency of different documents’ views as dependent variables rejects equality of countries at $p < 0.001$ but not equality of common and civil law ($p = 0.78$). However, ANOVA does not take into account the ordering of document views, which one might argue is the most characteristic aspect of possibly different ways of thinking. To capture this important dimension, we introduce sequence analysis to comparative law: we discretize all document paths to 500 steps of equal normalized length, calculate the Levenshtein distance between any two such sequences, aggregate the within-group distances normalized by total distances, and then compare this normalized aggregate to its empirical distribution under random permutation of group labels. In more detail,

---

5 Zweigert & Kötz (1998 §5.III.2), Legrand (1996, 1999), and others claim that civil law thinking is abstract, systematizing, and institution-focused while common law thinking is more concrete, casuistic, and fact-based. None of these differences seems sufficiently specifically articulated, however, for rigorous testing.
Levenshtein distance is a standard metric for distance between sequences of categorical data in computer science, linguistics, and other fields. It is the minimum number of edit operations (delete, insert, or substitute) required to transform one sequence into the other. (We obtain identical results with substitution cost equal to 1.5 times the cost of insert or delete. We use Halpin’s (2017) implementation.) Then letting $SS_T$ be the total sum of all pairwise Levenshtein distances divided by the total number of observations (299), and $SS_W$ be the sum of such sums calculated separately for each group, the pseudo-$R^2$ $(SS_T - SS_W)/SS_T$ is a well-behaved, useful measure of association of sequences and groups (Studer et al. 2011). A permutation test is exact even in finite samples against the randomization null hypothesis that there are no differences at all between groups (Lehmann & Romano 2005 Theorem 15.2.2).

When the groups are countries, the pseudo-$R^2$ equals 0.04. This is low, but much higher than random: in 100,000 random re-assignments of all country labels to judges, we not once obtained a pseudo-$R^2$ this high, i.e. $p \leq 10^{-5}$. In other words, there are detectable country differences in the document view paths.

When the groups are common and civil law, the pseudo-$R^2$ is much lower still (<0.01). To assess its statistical significance, we need to permute common/civil law labels of countries, not of individuals, because the common/civil law
hypothesis and corresponding null hypothesis is about the similarity of entire
countries and hence groups of judges, not merely of judges within countries.\textsuperscript{6} With two group members and seven countries, we reject equality of common and civil law at \( p \leq 0.05 \) only if the two common law countries India and USA generate the most extreme pseudo-\( R^2 \) of any of the \( \binom{7}{2} = 21 \) possible country pairs. While this seems to be a demanding test, the extreme test statistic is arguably exactly what one should expect if common and civil lawyers really thought differently. In any event, we have found empirically that the test is not so demanding after all, i.e. the test has high power even for relatively small differences. In simulated data drawn with replacement from the entire population of judges, even a 1 percent “half-time” sliver (the middle 5 out of the 500 steps) set to precedent for (arbitrarily designated) “common law judges” and to statute for the other, “civil law” judges is detected 72 percent of the time, and a 2 percent sliver is detected 99 percent of the time. When we also simulate within-country similarities by first drawing countries and then drawing judges only within countries, in each case with replacement, we still estimate 93 percent power for detecting a 5 percent sliver. In words, if there are meaningful differences between common and civil law, our permutation test has high power to detect them at \( p < 0.05 \). However, in actuality, our test does not come close to rejecting equality of common and civil law (\( p = 0.29 \), i.e. five other country pairs generate a higher pseudo-\( R^2 \)). The underlying reason is that judges from the two common law countries India and USA are very different. For example, U.S. and Indian judges spend, respectively, the lowest and second-highest average fraction of time with the statute, and the second highest and second lowest with the precedent. We return in the conclusion to the question of whether other common law countries might be more similar.

As supplementary comparative evidence, Figure 2 presents the country-prevalence of key arguments (statute, policy, and precedent) in the written reasons, which all but 9 of the 299 participants submitted and which are reproduced in Supplementary Appendix C (in expert English translation, where applicable). The upper panel presents the raw country-prevalence of each argument, i.e. the fraction of judges from that country that mentioned the argument in their reasons. The lower panel divides the country-prevalence by the

\textsuperscript{6} Intuitively, it is easy to see that individual permutation would be wrong: if all within-country distances were zero (i.e., all judges from one country were the same) and all countries were equidistant (i.e. no group of countries were special), then with probability approaching one, permuting individual labels would yield greater “within-group” distances and hence a lower pseudo-\( R^2 \) than any actual group of countries simply because the permutation leads to more mixing of judges from different countries.
average number of words written by judges from the respective country (in English translation, where applicable).\textsuperscript{7}

We again see considerable variation within countries and across countries (now even with the naked eye), but not between common and civil law. In fact, the biggest country differences are between the two common law countries India and USA, who occupy opposite extremes for the prevalence of precedent and, when scaled by average words, policy, and whose differences are larger than those between the common and civil law means on all six dimensions. As with document paths, permutation tests strongly reject equality of the three prevalences—scaled or not—in the seven countries ($p < 10^{-4}$) but not in common and civil law ($p = 0.57$); see Supplementary Information S5 for the technical details. To the extent written reasons are probative about judges’ thinking (see the general caveat in the ‘Introduction’ Section), they thus confirm our finding from document paths: judges “think differently” in different countries, but there is no evidence that these differences are related to the common/civil law distinction.

\textsuperscript{7} In the raw text data, average length of reasons measured in English words (mean 172, SD 140) differs considerably between countries. Chinese (mean 74) and U.S. (mean 94) judges tend to write short reasons, French (mean 262), and German (mean 227) judges long reasons, with Argentinian (mean 168), Brazilian (mean 165), and Indian (mean 155) judges in between. The differences in length might themselves be a difference in style, or simply reflect experimental artifacts such as the greater ease of writing on a laptop (which we used in France and Germany) compared to on an iPad (which we used in the USA and partly in China).
Notice that our last finding concerns the *arguments* in judges’ reasons and thus is not inconsistent with the common view reported in the Introduction that there are systematic differences in the *style* of judicial opinions. Expert readers can arguably detect such differences also in our judges’ written reasons. We have verified that natural language processing (NLP) algorithms can detect these differences too. NLP also shows, however, how superficial such differences can be: NLP can detect source language differences in translations (Rabinovich & Wintner 2015), but that alone would hardly be considered evidence that the speakers of different languages “think differently” (although they might). Applied to judicial reasons, NLP might simply recover the trivial fact that the common law’s language is English and the civil law’s is not, which is why we have not pursued this route.

**Effect of Law (Horizontal Precedent)**

We now turn to our second question: does horizontal precedent have an effect on judges’ decisions, or perhaps more to the point, how does its estimated effect size compare to judicial biases that have been documented in the literature? To emphasize that this is not a comparative question and to facilitate reading, Figure 3 plots our experimental evidence for all countries combined (country-specific plots are in Supplementary Information S4). The left panel plots affirmance rates by randomly assigned horizontal precedent and defendant, and the right panel plots sentence length by anchor.

All estimated effects go in the expected direction. The defendant’s conviction is affirmed more frequently under *Affirm* than under *reverse*, and more frequently under *reverse* than under *REVERSE*. The unsympathetic defendant’s conviction is affirmed more often than the sympathetic defendant’s. The distribution of sentence lengths shifts right as the anchor increases from ten to twenty-five to forty.

However, the estimated precedent effect is the smallest and the most likely to be mere noise. The affirmation rate under *REVERSE* (*reverse*) is only six (five) percentage points lower than under *Affirm*, whereas the difference between the two defendants is nine percentage points. Fisher’s exact two-sided *p*-values are 0.16 and 0.19 for the two precedent comparisons but 0.06 for the defendant comparison; the anchor-sentences rank-correlation $\rho = 0.16$ has

---

8 *REVERSE*’s confidence interval is larger than the other two precedents’ because we underweighted *REVERSE* in our randomization, see Supplementary Information S3.

9 The Fisher’s tests are stratified on country and the respective other factor (as in Jung 2014). This and the other results for precedent and defendant reported in this and the next paragraph are
$p < 0.01$ ($p = 2t(N-2, |\rho|(N-2)^{1/2}(1-\rho^2)^{-1/2})^{10}$). Perhaps, more to the point, the precedent effect’s upper 95% confidence bound is small in the absolute and relative to the defendant effect: The 95% confidence interval is [0.7–4.0] for the odds ratio (OR) of Affirm over REVERSE, and [0.3–2.6] for the ratio of the precedent and defendant odds ratios ($\text{OR(Affirm/REVERSE)}/\text{OR(Unsympathetic/Sympathetic)}$) (estimates from logit regression of affiliation on precedent and defendant conditional on country, critical values from normal distribution).

The precedent effect estimate is not meaningfully larger in subgroups that one might have expected to exhibit a stronger precedent effect. Judges who mention the precedent in their reasons exhibit twice the difference in affirmance rates between Affirm and REVERSE and between Affirm and reverse, but of course even the double differences are still small: eleven ($p = 0.14$) and ten ($p = 0.17$) percentage points, respectively. Judges from common law countries (which, recall, have stare decisis) do not exhibit stronger precedent effects; in fact, their observed behavior is counter to precedent: U.S. judges had higher affirmation rates under reverse than Affirm (REVERSE was not used in the USA, see note 4), and all Indian judges affirmed under REVERSE but not under Affirm and reverse.11

---

10 We obtain similar results in various regressions reported in Supplementary Information S5.

11 In contrast, Chinese judges analyzed separately do exhibit precedent effects (Liu et al. 2020).
DISCUSSION

Our results provide negative answers to both our questions: Common and civil law judges do not appear to think differently, and horizontal precedent does not affect judicial decisions, at least not much more than sympathy and anchoring biases. Naturally, we cannot exclude the existence of effects that are smaller than what we have power to detect, or that manifest only in different contexts. Still, our findings are important because the effects we study were supposed to be large and not restricted to particular contexts, and our 299 observations would have given us good power to detect even modest main effect sizes. We emphasize the comparison of precedent effect to biases for this reason: whether or not horizontal precedent ultimately has some effect, our troublesome finding is that the size of this effect is bounded to be not much larger than that of biases.

Specifically, regarding the common/civil law distinction, our finding is limited to our sample of countries, which for the common law includes only India and the USA. There might be more commonalities between common law countries of a similar level of development or in the Commonwealth. Still, it is important to know that any common law family resemblance does not extend to India and the USA. Moreover, our data also show that similarity cannot simply be assumed even between similarly developed, geographically adjacent countries of the same family; our document path and reason data show considerable differences between France and Germany (they only rank tenth in proximity among the twenty-one country pairs).

Continuing with the horizontal precedent effect, there are many real-world examples where precedent clearly does matter and does determine case outcomes. In functional court systems, lower courts respect vertical precedent set by higher courts, and each court (or panel of a court) tends to respect its own precedent as long as the court’s membership is stable. For example, ever since the U.S. Supreme Court ruled same-sex marriage constitutionally protected in Obergefell v. Hodges (2015), every U.S. court has respected this precedent, including the U.S. Supreme Court itself and lower courts that had previously decided differently. Such respect is overdetermined, however, by the threat of reversal for lower courts and persistent judicial preferences on the higher court (unless its composition changes significantly). When judges operate outside of a hierarchy—specifically, federal judges deciding a state law question—Klerman & Spamann (2019) barely find an effect even of vertical precedent, i.e. precedent set by a court of nominally higher authority (namely the State Supreme Court). The observable respect of precedent in real-world situations such as Obergefell’s reception is also predicated on an active shared
understanding of what the precedent stands for, which is often absent when judges refer to precedent. For example, *Obergefell v. Hodges* itself cited over 100 precedents by various courts, some over 100 years old, none of which was directly on point, and many of them were thus subject to divergent characterizations by the judges and their audience. This is also typical when lower courts address a novel legal issue. In such cases, our results support the suspicion (e.g. Schauer 2018) that precedent has little effect on decisions, whatever its rhetorical appeal.

To be sure, our design addresses only one of many constellations in which precedent might matter. In addition, the realism of our design and that of Klerman & Spamann (2019) is imperfect. Most importantly, our judges decide anonymously, whereas real-world judgments are signed and hence implicate reputational consequences, including career concerns for lower judges. We believe that the most pressing agenda for future research on judicial decision-making is to investigate this and similar ways and areas in which law might matter for judicial decisions, and how this may differ across jurisdictions. Our findings and those of Klerman & Spamann (2019) show law to be less determinative than previously thought (cf. Spamann & Klöhn 2016 §3). But counterexamples such as *Obergefell*’s reception show that legal nihilism is equally inapposite. Tracing the boundaries between these and similar situations, and identifying cross-country differences in these boundaries, will help us understand and improve law’s important role in society.

Similarly, our finding of considerable country differences coupled with the inability of the common/civil law distinction to explain any of them suggests a fruitful search for alternative classifications of legal systems, which is already under way (e.g. Glenn 2014). We hope that the document path evidence we introduce in this paper can be of great help in this search. These novel data can presumably be analyzed in innovative ways that we have not even thought of.

**SUPPLEMENTARY MATERIAL**

Supplementary material is available at JLA online.

**REFERENCES**

Chung, EunYi & Joseph Romano, P. 2016. Multivariate and Multiple Permutation Tests. 193 *J. Econ.* 76–91.

Englich, Birte & Thomas Mussweiler. 2001. Sentencing under Uncertainty: Anchoring Effects in the Courtroom. 31 *J. Appl. Soc. Psychol.* 1535–1551.
Englich, Birte, Thomas Mussweiler & Fritz Strack. 2006. Playing Dice with Criminal Sentences: The Influence of Irrelevant Anchors on Experts’ Judicial Decision Making. 32 Pers. Soc. Psychol. Bull. 188–200.

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

Fischman, Joshua B. 2014. Measuring Inconsistency, Indeterminacy, and Error in Adjudication. 16 Am. Law Econ. Rev. 40–85.

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

Gény, François. 1899. Méthode d’Interprétation et Sources en Droit Privé Positif. Paris: L.G.D.J.

Glenn, H. Patrick. 2014. Legal Traditions of the World: Sustainable Diversity in Law. 5th Edition. Oxford: Oxford University Press.

Guthrie, Chris, Jeffrey J. Rachlinski & Andrew J. Wistrich. 2001. Inside the Judicial Mind. 86 Cornell Law Rev. 777–830.

——— 2007. Blinking on the Bench. 93 Cornell Law Rev. 1–43.

Halpin, Brendan. 2017. SADI: Sequence Analysis Tools for Stata. 17 Stata J. 546–572.

Harris, Allison P. & Maya Sen. 2019. Bias and Judging. 22 Annu. Rev. Polit. Sci. 241–259.

Henderson, M. Todd & William H. J. Hubbard. 2015. Judicial Noncompliance with Mandatory Procedural Rules under the Private Securities Litigation Reform Act. 44 J. Legal Stud. S87–S105.

Jung, Sin-Ho. 2014. Stratified Fisher’s Exact Test and Its Sample Size Calculation. 56 Biometric. J. 129–140.

Kahan, Dan M. 2015. Laws of Cognition and the Cognition of Law. 135 Cognition 56–60.

Kahan, Dan M., David Hoffman, Danieli Evans, Neal Devins, Eugene Lucci & Katherine Cheng. 2016. “Ideology” or “Situation Sense”? An Experimental Investigation of Motivated Reasoning and Professional Judgment. 164 Univ. Pennsylvania Law Rev. 349–439.

Kantorowicz, Hermann [Gnaeus Flavius]. 1906. Der Kampf um die Rechtswissenschaft. Heidelberg: Carl Winter’s Universitätsbuchhandlung.

Kennedy, Duncan. 1998. A Critique of Adjudication (fin de siècle). Cambridge, MA: Harvard University Press.

Klerman, Daniel & Holger Spamann. 2019. Law Matters – Less Than We Thought. SSRN Working Paper. https://ssrn.com/abstract=3439526.

Kötz, Hein. 1982. Die Begründung Höchstrichterlicher Urteile. In: Nederlandse Vereniging voor Rechtsvergelijking. Preadviezen Uitgebracht
Voor de Nederlandse Vereniging Voor Rechtsvergelijking, vol. 32, 5–21. Deventer: Kluwer.

La Porta, Rafael, Florencio Lopez-de-Silanes & Andrei Shleifer. 2008. The Economic Consequences of Legal Origins. 46 J. Econ. Lit 285–332.

Lasser, Mitchel de S.-O.-L’E. 2004. Judicial Deliberations: A Comparative Analysis of Judicial Transparency and Legitimacy. New York: Oxford University Press.

Legrand, Pierre. 1996. European Legal Systems Are Not Converging. 45 Int. Comp. Law Quart. 52–81.

——— 1999. Fragments on Law-as-Culture. Deventer: W.E.J. Tjeenk Willink.

Lehmann, Erich L. & Joseph P. Romano. 2005. Testing Statistical Hypotheses. New York: Springer.

Liu, John Zhuang, Lars Klöhn & Holger Spamann. 2020. Precedent and Chinese Judges: An Experiment. Am. J. Comp. Law.

Llewellyn, Karl N. 1930. A Realistic Jurisprudence—the Next Step. 30 Columbia Law Review 431–465.

Merryman, John & Rogelio Pérez-Perdomo. 2007. The Civil Law Tradition. 3rd Edition. Stanford: Stanford University Press.

Nadler, Janice & Mary-Hunter McDonell. 2012. Moral Character, Motive, and the Psychology of Blame. 97 Cornell Law Rev. 255–304.

Rabinovich, Ella & Shuly Wintner. 2015. Unsupervised Identification of Translationese. 3 Trans. Assoc. Comput. Linguist. 419–432.

Rachlinski, Jeffrey J. & Andrew J. Wistrich. 2017. Judging the Judiciary by the Numbers: Empirical Research on Judges. 13 Annu. Rev. Law Soc. Sci. 203–229.

Rachlinski, Jeffrey J., Andrew J. Wistrich & Chris Guthrie. 2015. Can Judges Make Reliable Numeric Judgments? Distorted Damages and Skewed Sentences. 90 Indiana Law J. 695–739.

Rachlinski, Jeffrey J., Sheri Johnson, Andrew J. Wistrich & Chris Guthrie. 2009. Does Unconscious Racial Bias Affect Trial Judges? 84 Notre Dame Law Rev. 1195–1246.

Schauer, Frederick. 2018. Stare Decisis—Rhetoric and Reality in the Supreme Court. 2018 Supreme Court Rev. 121–143.

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

Spamann, Holger. 2015. Empirical Comparative Law. 11 Annu. Rev. Law Soc. Sci. 131–53.

Spamann, Holger & Lars Klöhn. 2016. Justice is Less Blind, and Less Legalistic, than We Thought: Evidence from an Experiment with Real Judges. 45 J. Legal Stud. 255–280.
Spellman, Barbara & Frederick Schauer. 2012. Legal Reasoning. In Holyoak K. J. & R. G. Morrison, eds., *The Oxford Handbook of Thinking and Reasoning*. 2nd Edition. Oxford: Oxford University Press.

Studer, Matthias, Gilbert Ritschard, Alexis Gabadinho & Nicolas S. Müller. 2011. Discrepancy Analysis of State Sequences. 40 *Sociol. Methods Res.* 471–510.

Summers, Robert S. & Michele Taruffo. 1991. Interpretation and Comparative Analysis. In Neil MacCormick D. & S. Summers Robert, eds., *Interpreting Statutes: A Comparative Study*. Brookfield, VT: Dartmouth.

Wetter, J. Gillis. 1960. *The Styles of Appellate Judicial Opinions: A Case Study in Comparative Law*. Leyden: A. W. Sythoff.

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

Zweigert, Konrad & Hein Kötz. 1998. *An Introduction to Comparative Law (Tony Weir Transl).* 3rd Edition. New York: Oxford University Press.