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

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

CAN LAW STUDENTS REPLACE JUDGES IN EXPERIMENTS OF JUDICIAL DECISION-MAKING?

Holger Spamann
Lars Klöhn

Discussion Paper No. 1093

02/2023

Harvard Law School
Cambridge, MA 02138

Can Law Students Replace Judges in Experiments of Judicial Decision-Making?

Holger Spamann[*] **Lars Klöhn**[♠]

Abstract: Experimental research on judicial decision-making is hampered by the difficulty of recruiting judges as experimental participants. Can students be used in judges' stead? Unfortunately, no. We ran the same high-context 2×2 factorial experiment of judicial decision-making focused on legal reasoning with 31 U.S. federal judges and 91 elite U.S. law students. We obtained diametrically opposed results. Judges' decisions were strongly associated with one factor (sympathy, i.e., bias) but not the other (law). For students, it was the other way around. Equality between the two groups is strongly rejected. Equality of document-view patterns—a proxy for thought processes—and written reasons is also strongly rejected.

3/24/2023

[*] Lawrence R. Grove Professor of law, Harvard Law School. hspamann@law.harvard.edu.

[♠] Professor of law, Humboldt-University Berlin. lars.kloehn@rewi.hu-berlin.de.

We thank Ludwig-Maximilians-Universität for financial support; Roland Ramthun of Documstorm for programming the experimental interface; Julian Franke, JohneHenre Rzodkiewicz, and Oliver Schön for research assistance; Alex Whiting for suggesting the aiding and abetting controversy; Nancy Gertner, John Manning, Denise Neary, and the FJC for arranging the session at which the judge experiment was conducted; the three law school deans for permission to advertise to their students; and above all the judges and students for their participation. For very helpful comments, we thank seminar participants at Berkeley, Harvard, UT Austin, UVA, the Conference on Empirical Legal Studies 2022 (especially our commentator Greg Mitchell), the Experimental Methods in Legal Scholarship Conference 2022 (especially our commentator Alex Morell), the Michigan-USC-Virginia Virtual Law and Economics Series, and Michael Gilbert. This research was approved by Harvard University's Committee on the Use of Human Subjects as Protocol # IRB15-0206. All data and code are available at doi.org/10.7910/DVN/3SRIDI with the exception of raw U.S. judge demographics, which pursuant to IRB conditions can only be obtained from Spamann upon signing a confidentiality agreement. All experimental materials are included as appendices to this paper or Spamann & Klöhn (2016).

1 Introduction

Experiments are the gold-standard of causal inference, and the causal determinants of judicial decisions are of obvious interest. Nevertheless, experiments with judges are rare. Judges are few and busy. The organizations that could mobilize them in greater number—courts, judges' associations, and judicial agencies like the FJC—are wary of doing so.[1] Some scholars have succeeded to recruit sufficient numbers of judges for vignette experiments over several rounds of judicial conferences or continuing legal education seminars.[2] Only two studies—one of which we replicate with students—have recruited judges for longer experiments mimicking features of real-world judicial decision-making.[3] In judges' stead, many studies of legal reasoning and judicial decision-making employ (law) students.[4] If students were good proxies for judges, the rate of scientific discovery could be greatly enhanced. Are they?

Unfortunately, this paper's answer is no. We conducted the same 2×2 factorial between-subject experiment with 31 U.S. federal judges and with 91 law students at three top-ranked U.S. law schools. Specifically, participants spend up to an hour deciding a fully briefed appeal in an international war crimes case. The experimental variations are (1) whether the precedent favors or disfavors the defendant and (2) whether the defendant is sympathetic or unsympathetic (in legally irrelevant ways). We separately reported the judge results in Spamann & Klöhn (2016): the 31 judges disproportionately ruled in favor of the more sympathetic defendant but were unmoved by precedent. As we report now, the 91 students did the opposite: their decisions did not differ between defendants but did differ strongly between precedents. In short, if one had run Spamann & Klöhn (2016) with students instead of judges, one would have found the opposite results. The probability of observing such differences by chance—if judge and student populations did not differ—is estimated to be only one in 500, i.e., the null hypothesis of equality of the two effects in the two populations is rejected at *p*<0.002. Beyond their ultimate decisions, we also document that students significantly differ from judges in the reasons they write and in their reasoning process, to the extent we can observe it by tracking their use of legal documents in the experiment.

Our experiment gives students their best shot at mirroring judges' behavior. We recruited our student participants at the very kind of schools—top-ranked U.S. law schools—that supply a large share of the U.S. federal judge population from which we draw our comparison sample (Iuliano & Stewart 2016).[5] (Perhaps some other type of judge than U.S. federal judges can be better proxied by students, but we would not know why.[6]) While our student and judge samples are not randomly drawn from their respective populations—participation is voluntary—we have no reason to suspect selection related to the results.[7] As already mentioned, the usual tools of statistical inference show that our findings are very unlikely to be due to chance. Finally, our experimental task, while relatively realistic, is one that students

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

[2] E.g., Rachlinski et al. (2009, 2015); Wistrich et al. (2014); Kahan et al. (2016).

[3] Spamann & Klöhn (2016) (also conducted in China (Liu, Klöhn, & Spamann (2021)) and five other countries (Spamann et al. (2021))); Klerman & Spamann (2024).

[4] E.g., Braman & Nelson (2007); Furgeson et al. (2008a, 2008b); Feess & Sarel (2018); Engel & Grgić-Hlača (2021).

[5] To be sure, bankruptcy judges and magistrates, which are also in our sample, may hail from a more diverse set of schools.

[6] At a minimum, our results show that the behavior of U.S. federal judges—who attract substantial scholarly and popular attention—cannot be proxied by law students.

[7] Representative samples of students would not even be helpful because the average student does not become a judge. Other experimental studies with students do not randomly sample from the student population either.

can easily comprehend. Realism may be required for ecological validity, i.e., for experimental results to generalize to judges' behavior in the real world (Holste & Spamann, forthcoming). Many real-world judicial tasks are either too specialized for students to understand (e.g., complex procedural motions) or too routinized to think students could possibly do it the same way (e.g., bail decisions), or both. Our task confronted participants with a legal question that was easy to understand but novel for all (which is not unusual for federal judges, who have a varied, generalist docket involving the laws of various states). To be sure, it is theoretically possible that there is something idiosyncratic about our experimental task that divides judges and students in a way that most judicial tasks would not, but we would not know what that would be.[8]

We mostly resist the temptation to speculate *why* judges are different. This question cannot be convincingly answered with our data, if at all, and even if it could, the answer would be of limited practical use. Judges and students differ on too many dimensions. Besides the status, experience, etc. of being a judge itself, judges have more and different pre-appointment experience than students, and they are selected from the much larger pool of (former) law students. Much of the selection is on unobservable dimensions (what type of person wants to become a judge?). In any event, our data only contain the most rudimentary demographic information, and our sample is too small to disentangle multiple variables. From the practical perspective of designing valid experiments of judicial decision-making, the explanation would be useful only if it identified a subset of students that can stand in for judges. For explanations such as age, that would not be the case because virtually all students are much younger. The more practically relevant question is if one can identify *another* subject pool, such as practicing attorneys, that behaves like judges in experiments; this is a pressing question for future research. Of course, research should ultimately develop a unified model of decision-making that explains how and why different types of subjects behave differently. But we are far away from that level of understanding.

Our findings are consistent with the expertise literature and with the majority of papers that have conducted identical experiments with students as with judges, albeit with less realism and without our methodological focus. The general expertise literature tends to find that some experts perform reliably differently than non-experts, albeit only within a very narrow domain of expertise (Ericsson 2018). Judges have years or even decades of training and experience in legal analysis and are selected specifically for this skill. They are thought to behave differently specifically for *legal* decisions (Llewellyn 1940; Schauer 2010; Kahan et al. 2016; Spamann & Klöhn 2016). How the difference manifests in any given application is a priori unclear: judges might be particularly able to ignore non-legal influences, or particularly able to manipulate legal materials to get to their desired result (Kennedy 1998). Our finding is consistent with the latter, while Redding & Reppucci (1999) and Kahan et al. (2016) are consistent with the former, but all three studies find that judges and students behave differently. The only experimental legal reasoning study that does not find differences between judges and students is Chen & Li (2018). Neither of these three other experiments was as realistic as ours, which may have dampened differences between judges and students. Many but not all studies of fact-finding also find differences between judges and students. By contrast, judges and students unsurprisingly perform similarly on tasks that are not specific to judging. Holste & Spamann (forthcoming) provide a recent thorough review.

[8] It is also theoretically possible that students in the lab behave like judges in the real world—the ultimate behavior of interest—while judges themselves do not. We find implausible such "non-monotonicity" of ecological validity as a function of realism.

We do not argue that students cannot have any role in the study of judicial decision-making. For example, Gilbert (2011) uses students' survey answers on interpretive questions as a legal baseline against which to compare real-world judicial decisions. Our point is that the students cannot stand in for the judges themselves. Gilbert's finding that students' interpretations only partially predict judicial decisions is consistent with this point.

The rest of this paper is structured as follows. Section 2 explains the experimental design. Section 3 describes the sample. Section 4 reports the results. Section 5 concludes. All data and code are available at doi.org/10.7910/DVN/3SRIDI, and all experimental materials are included as appendices here or in Spamann & Klöhn (2016).

2 Experimental Design

The main point of the present paper is, of course, to run the same experiment in two different samples that we describe in the subsequent section 3. The experiment itself is described in detail in Spamann & Klöhn (2016); here we only give a brief summary. All the study materials are provided in the online appendix of Spamann & Klöhn (2016), except the materials that differed for students: the recruitment ad reproduced in section 3.2 below, and the student informed consent form and exit questionnaires reproduced in this paper's Appendix.

The experiment asks participants to imagine being a judge on the appeals chamber of the International Criminal Court for the Former Yugoslavia (ICTY) deciding a defendant's appeal of his conviction for war crimes by the ICTY's trial chamber. Participants have 50 minutes to decide. Participants receive a statement of agreed facts (written by us), briefs for the defendant and the prosecution (also written by us), the judgment below (taken from a real case, *Prosecutor v. Perišić*[9], with only names and dates altered), a precedent (discussed below), and the ICTY statute. Besides their binary affirm/reverse decision, participants are asked to indicate brief reasons for their decision in a text box (on the same page), and to indicate an appropriate sentence on the next page.[10]

Participants are randomly assigned one of two defendants and one of two precedents.[11] The randomization mechanism was designed to create 2×2 cells of equal sample size. In the judge sample, technical difficulties lead to slight group imbalance, as described in Spamann & Klöhn (2016).

Both defendants were fictitious military chiefs in Croatia and Serbia, respectively, responsible for organizing logistical support to their respective ethnic brethren in the Bosnian civil war. Besides the arguably different valence of "Croat" and "Serb" in U.S. perceptions of this war (NATO eventually bombed Serb positions, also again in Kosovo several years later), we interspersed several positive post-war facts, such as remorse, about the Croat and negative post-war facts about the Serb, such as spite for the ICTY, in the respective statement of facts, the trial judgment facts, and the briefs. We chose these attributes

[9] Prosecutor v. Momčilo Perišić, Case No. IT-04-81-T, judgment of the trial chamber of 9/6/2011.

[10] The sentencing task was embedded in a separate anchoring manipulation. We do not use this information here or in Spamann & Klöhn (2016), as the anchoring test has sufficient power only in the large sample of Spamann et al. (2021).

[11] As a pilot for Spamann et al. (2021), nineteen student recruits—not included in our 91 student sample—were assigned a third precedent that even more strongly favored reversal. The results for these students are consistent with those reported here (the precedent appears to have a strong effect in the expected direction) but not reported because they cannot be compared to any judge participants with the same precedent.

and their depiction to be clearly irrelevant from a strictly legal perspective, at least for the decision at hand (determination of guilt, not sentencing). We named the fictitious Croat and Serb defendants Horvat and Vuković, respectively, but for ease of reference this paper refers to them simply as "sympathetic" and "unsympathetic." For a fuller description, see Spamann & Klöhn (2016) and the full experimental materials in its online appendix.

One of the two precedents weakly favors the defendant's position (*obiter dictum*), whereas the other weakly disfavors it (based on distinguishable facts). The briefs are adjusted accordingly. Specifically, the main question in *Perišić*—and the only question in our stripped-down case—was whether the defendant's logistical support to the foreign group qualified as aiding and abetting under Article 7(1) of the ICTY statute even though it was not "specifically directed" at the war crimes (as opposed to the general war efforts). (There was no doubt that the defendant knew that the group committed war crimes.) The Trial Chamber in *Perišić*—and thus in our case—had answered in the affirmative. One of our two precedents supported this position: the ICTY Appeals Chamber's *Šainović* opinion held "[t]hat 'specific direction' is not an element of aiding and abetting liability under customary international law" and affirmed the conviction of a defendant who did not "specifically direct" his aid at the crimes.[12] By contrast, our second precedent, the ICTY Appeals Chamber's *Vasiljević* opinion, provides weak support for the contrary conclusion, defining aiding and abetting *obiter dictum* as "*specifically directed* to assist, encourage or lend moral support to the perpetration of a certain specific crime."[13] For ease of reference and as a mnemonic for their holdings' implications for our case, this paper refers to *Vasiljević* and *Šainović* as *reverse* and *Affirm*, respectively. Each participant receives a package with only one of *reverse* and *Affirm*, and with briefs discussing only that one precedent.

Participants perform the task on an iPad or computer as described in section 3. The machine records the paragraphs of the materials active on a participant's screen in 10-second intervals, which we use to construct and compare "document view paths" as described in section 4.2.

3 Samples

Table 1 summarizes our judge and student samples. We could not collect much demographic information because of concerns about identifiability in the small judge sample and heightened sensitivity in this population.[14] We did, however, ask about gender and 10-year age groups (although not for students since that seemed redundant). We also asked a few exit questions and measured the time spent with the documents.

[12] Prosecutor v. Sainović, Case No. IT-05-87-A, at para. 1649.

[13] Prosecutor v. Vasiljević, Case No. IT-98-32-A, at para. 102 (emphasis added).

[14] We do have the confidential list of 47 judges invited to the workshop and have collected their demographics. But the median age of the 47 invitees is a decade higher than the median age of the 31 participants, which would render suspect any inference from the 47 invitees to the 31 participants on other publicly observable characteristics, such as appointing president for Article III judges.

Table 1: Summary Statistics

Participant type	judge	student
N	31	91
with reasons	29	91
with sentence	28	90
Female (mean)	0.18	0.48
Age group (median)	55-65	
Fraction 3Ls (vs. 2Ls) (student round 2)		0.53
Prior knowledge of international criminal law (mean)	0.07	0.10
Recognized names and places in the case (mean)	0.82	0.51
Confidence in decision (mean)	0.65	0.62
Minutes spent with documents (mean)	36	27
Affirmed (mean)	0.74	0.77
Sentence (median in years)	25	10

We did not conduct dedicated comprehension checks. We did, however, ask two research assistants independently to read all judgment reasons and flag any that indicated a misunderstanding of the task.[15] The coding protocol is provided in Appendix A.4. Neither assistant found such a case.

3.1 Judges

We conducted the judge round of the experiment at an annual three-day workshop for U.S. federal judges organized jointly by Harvard Law School and the Federal Judicial Center in April 2015. All participants were U.S. federal judges including circuit judges, district judges, bankruptcy judges, and magistrates.[16] About 50 different judges attend each year, implying that a sizeable fraction of the approximately 1,800 federal Article III, bankruptcy, and magistrate judges must attend over time, such that selection into attendance cannot be too skewed.[17]

The experiment was part of a session on "Behavioral Research on Judicial Decision-Making" in the middle of the second morning. Several weeks earlier, the judges had received an invitation to the experiment with all consent-relevant information (Spamann & Klöhn 2016, online appendix A.1.1) and a reading "assignment": Guthrie et al. (2007), which discusses biases in judicial fact-finding. The experiment was administered on iPads we provided to the judges. Participation was voluntary—informed consent was obtained on the first screen—but all the judges present in the room participated in the experiment. We

[15] We also asked the research assistants if a participant verbally indicated that he or she did not want to submit a judgment (even though the system recorded one, perhaps because of data entry error). As explained in 18, we had already excluded from the sample the one judge who had merely written "Not sufficient time to form a judgment." The assistants only flagged two other judges who began their reasons with "While insufficient time has been allowed to analyze all aspects of this case …" or "Although the time constraints of this exercise limited my ability to master the facts …," respectively, but proceeded to give full reasons; these remain in the data.

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

[17] We estimate the total number of federal judges from tables 10-12 of https://www.uscourts.gov/statistics-reports/judicial-business-2021.

lost some small number of participants due to technical problems with the iPads described in Spamann & Klöhn (2016). We ultimately have 31 judge participants.[18]

3.2 Students

With permission of the respective deans, we recruited students at three top-ten U.S. law schools with the following ad sent through schools' email lists and, in one case, Facebook groups used for announcing events and job opportunities:

> Title of ad: "Get paid to learn some international criminal law!"
>
> Text of ad: "If you are a 2L or 3L who has NOT taken either of international criminal law or international humanitarian law, you can participate in an online study conducted by [X]. You will spend an hour judging a fictitious but highly realistic and highly topical appeals case in that area. All the legal materials you need will be provided to you online. Upon completion of the task, you will receive an Amazon gift code for $20. You can access more information on the task here [link to the consent form at the online site of the experiment]."

In this quote, we have obscured the name, title, and affiliation of the faculty member X to avoid identification of the schools involved, as requested by their deans.

[18] The number of judge participants reported here is one lower—the net effect of dropping two and adding one—and three observations are recorded with the opposite outcome than in Spamann & Klöhn (2016). The reason is that our data cleaning script for this paper prioritizes data integrity to ensure consistency between judges and students, whereas Spamann & Klöhn (2016) had tied its hands with a hands-off approach (while reporting in footnotes 10 and 13 various robustness checks showing these potential issues did not matter). Nothing ultimately hinges on these changes, as we obtain only slightly weaker results with the judge data of Spamann & Klöhn (2016) (e.g., the paper's most important *p*-value, that from the joint score test in model 3 of table 3, would be 0.013 instead of 0.002). Our publicly posted data ingestion script compares the two datasets and identifies the individual observations that differ, and our publicly posted main analysis script presents the full analysis using either judge data set.

The underlying issue is that we fixed threats to data integrity in the backend of our experimental software after we collected the judge data and about 20% of the student data (student round 1, see 3.2 below) but before collecting the remaining student data (student round 2). Our original software had recorded both reversal and missing judgment as zero and, as discussed in Spamann & Klöhn (2016, n. 10), froze some iPads that RAs in the room swapped for new ones without ensuring that the new ones were set to the same 2×2 treatment combination. The judges also had to complete the study on iPads (i.e., with only a touchscreen and no mouse), which several of them reported to be unusual and uncomfortable and thus presumably prone to data entry error, whereas students could complete the study on a device of their choice. The judge data would thus likely contain systematically more errors if we followed a hands-off approach.

Therefore, we cross-check all participants' judgments against their reasons and correct the former to that implied by the latter in the five cases (each a judge) of a clear discrepancy. This includes one judge that we exclude because the reasons "Not sufficient time to form a judgment" suggest the judgment "0" is a missing value, and one judge that we add because while the participation did not advance past the judgment stage (the reason for mechanical exclusion in Spamann & Klöhn (2016)), the presence of coherent written reasons indicates the judgment "0" was a reversal. We exclude one judge observation that timestamps clearly show to be a replacement iPad (activated 25 minutes after all others and only spent 7 seconds with the documents), as those had a ¾ chance of mismatch of recorded treatment combination vs. treatment combination seen by the participant. See our ingestion script for more details.

We insisted that student participants "NOT [have] taken either of international criminal law or international humanitarian law" to put them on equal footing with the judges, who generally do not have such cases on their docket and probably never took such class, certainly not recently. (As mentioned in the introduction, the fact that neither judges nor students are experts in the case's subject matter is a feature, not a bug.)

The ad and study ran first at school 1 in April 2015. We refer to this as student round 1, which recruited 21 participants. The ad ran again at schools 1 and 2 in late November 2015, at school 3 in early January 2016, and at schools 1 and 3 in early February 2016. We refer to this as student round 2, which recruited 70 participants (net of the three withdrawals mentioned below). The experiment site closed in mid-February 2016, long before Spamann & Klöhn (2016) was first publicly posted on SSRN.

In student round 2, our IRB requested changes to the informed consent form we had used in student round 1. These are marked in Appendix A.1. In student round 2, our IRB also required us to allow participants to withdraw their participation after completing the study and receiving the debriefing because our IRB then considered incomplete our pre-study description of our research goal as "to learn about the process of legal reasoning and the role of various legal materials therein." Three students exercised this right.

As mentioned in the ad, we restricted participation to 2Ls and 3Ls at the respective schools. We could not directly enforce this, but it would have been pointless—or at least unpaid—for others to participate, and only three did. The consent form reproduced in Appendix A.1 informed participants that they would have to provide their official school email address to receive the Amazon voucher. All participants requested the voucher. One participant did not provide an official law school email, and two provided emails identifying them as LLM students. Neither of these three received the voucher. That said, we could not exclude these three participations because to protect anonymity of experimental responses, emails were collected in a separate file without a cross-walk to the experimental responses.

3.3 Differences between Judge and Student Samples

There were only minor differences between the judge and student versions of the experiments. While any difference in setup might theoretically explain differences in results, we do not think this is plausible. If it were, then experimental research would be in even deeper trouble than student/judge differences: if the minor differences catalogued below impacted results enough to explain a non-negligeable part of our student-judge effect differences, no lab experiment could generate useful information about the real world. Note in this respect that the differences catalogued here are not differences in experimental *treatment* – within each subject group (judges or students), treatment and control group were completely statistically identical except for the explicit experimental variation. Thus, the differences cannot confound treatment results within groups. The only question is whether they vitiate comparison between groups. That is a question akin to ecological validity. Our argument is that if one thought the differences in administration here were large enough to explain such massive differences in results, then to be consistent, one should deny ecological validity of virtually any lab experiment, and vice versa.

The only difference in the administration of the experiment is that all judges did the study at the same time in the same seminar room on an iPad that we provided, whereas students did the study at their leisure in a location of their choosing on their preferred device. Recall that judges and students were also (inevitably) recruited differently, and only students were paid (note that, in the real world, judges are not paid by the case). The difference in recruitment includes, for example, that judges already sat through the first day of the workshop before participating, whereas students did not.

The main experimental materials—between the instructions and the exit questions—were identical for judges and students.[19] These main materials are reproduced in the online appendix of Spamann & Klöhn (2016). The instructions (Appendix A.2) were also identical for judges and students except for two logistical adjustments relating to online vs. in-person administration of the study. The exit questionnaire for students (Appendix A.3) was adjusted to include more relevant questions (e.g., class year instead of whether they had previously worked as a prosecutor or public defender). The only documents that were more different for students and judges were the consent documents. When the students clicked on the link in the ad to participate in the study, they were first taken to the informed consent form reproduced in Appendix A.1. By contrast, the first screen on the judges' iPads was a short reminder of a letter sent before the session.[20] Together, the judges' letter and reminder contained the same information as that given to the students, albeit in abbreviated form.

As we emphasized in the introduction, we cannot determine *why* judges and students behave differently, in part because there are many plausible differences between the two populations that we do not observe. Here we comment only on the limited demographic information that we do have and show in Table 1. The differences between judges and students that stand out are that judges are obviously much older than most students, and that far fewer judges (0.18) than students (0.48) identified as female. Both were to be expected: in 2016, only 34% of U.S. federal judges identified as female, and in 2017, the median U.S. district judge was 61 years old.[21] In principle, these demographic differences could explain the differences in experimental behavior. We can rule this out empirically for gender since female and male students' decisions are indistinguishable (unreported). For age, we have no way of investigating the confound empirically even in principle—almost all students are young—but for that very same reason the confound is not practically relevant: even if judges and students differ "merely" because of age, students cannot replace judges in experiments because students are too young. The same practical irrelevance holds for a potential experience confound, perhaps proxied here by the higher percentage of judges (82%) than students (51%) who recognized names and places in the case. Of final note, judges spent 4/3 as much time with the documents as students.

We did, however, fix one typographical error in student round 2 that had been in student round 1 and the judges' round, which was that the *Šainović* variant of the prosecution's response brief contained an unintentional typo ("re*quir*ed" instead of "re*ject*ed") in its second out of three references to the precedent's holding. Judging by their written reasons, however, none of the judges and students in round 1 were misled by this.

[20] The letter is reproduced in the Online Appendix of Spamann & Klöhn (2016). The text shown on judges' first screen was:

> Thank you for participating in this study. I trust you have received my letter explaining the nature and purpose of the study; if not, you can read the letter here [hyperlink to letter].
> As you know, your participation is entirely voluntary, and you can discontinue the study at any point without penalty. To maintain the integrity of the research, I cannot answer questions during the study, but I am happy to answer any questions afterwards, either during debriefing or later.
> [continue button leading to the next page]

[21] Gender: Administrative Office of the U.S. Courts, The Judiciary Fair Employment Practices Annual Report: Fiscal Year 2019, Table 1 (FY 2016 numbers). Age: Congressional Research Service (Barry J. McMillion), U.S. Circuit and District Court Judges: Profile of Select Characteristics, August 1, 2017 (p. 23).

4 Results

4.1 Experimental Treatment Effects

Table 2 summarizes participants' decisions by treatment condition. Judges are in the left half of the table, students on the right. Within each group, the precedent treatment varies along the horizontal axis, and the defendant treatment along the vertical axis. For each of the four defendant × precedent combinations and the respective marginals, the table shows the share of the respective participants that upheld the conviction in decimal form and, in parentheses, as the fraction of affirmances over cell size (i.e., participants). (Recall that each participant only received one of the two precedents and judged only one of the two defendants.)

Table 2: Fraction Affirmed

		Judges			Students		
		Precedent			Precedent		
		Affirm	*reverse*	Total	*Affirm*	*reverse*	Total
Defendant	Unsympathetic	1.00 (7/7)	0.88 (7/8)	0.93 (14/15)	0.92 (23/25)	0.61 (14/23)	0.77 (37/48)
	Sympathetic	0.40 (4/10)	0.83 (5/6)	0.56 (9/16)	0.91 (21/23)	0.60 (12/20)	0.77 (33/43)
	Total	0.65 (11/17)	0.86 (12/14)	0.74 (23/31)	0.92 (44/48)	0.60 (26/43)	0.77 (77/91)

Each cell shows the fraction of participants that decided to affirm the defendant's conviction both as a single number and, in parentheses, as the fraction of affirmances over cell sample size

Focusing first on the left side of the table, and specifically on the bottom total row, it is readily apparent that the precedent made no difference for the judges. More judges affirmed the defendant's conviction under *reverse* (86%) than under *Affirm* (65%) – the opposite of the expected precedent effect. By contrast, as predicted, far more judges affirmed the conviction of the unsympathetic defendant (93%) than of the sympathetic defendant (56%) (Boschloo[22] two-sided p=0.024). These are the main results reported in Spamann & Klöhn (2016).[23]

The results for students on the right side of the table are the inverse. Students affirmed the conviction of the sympathetic and unsympathetic defendants in identical proportions (77%). By contrast, many more

[22] The Boschloo unconditional exact test is the recommended conceptually superior, uniformly more powerful generalization of the better-known conditional Fisher exact test (Mehrotra et al. 2003). We use the implementation of the test at https://www4.stat.ncsu.edu/~boos/exact/. That said, Fisher exact tests yield very similar results with our data, and we use them for the bootstrap tests below for simplicity.

[23] There are some small, immaterial differences between the results presented here and those in Spamann & Klöhn (2016) owing to the differences in data cleaning discussed in note 18.

students affirmed the conviction under the *Affirm* precedent (92%) than under *reverse* (60%) (Boschloo two-sided *p*=0.0004).

To test the difference in effects between judges and students formally, we estimate them in a joint regression. Table 3 shows a simple linear probability model (OLS, model 1) and a logit model estimated with conventional maximum likelihood (logit, model 2) or using the conditional distribution of the parameter sufficient statistics (exact logistic, model 3). In each case, the dependent variable is whether the participant affirmed the conviction, and the regressors are participant type, precedent, defendant, and interactions of participant type with precedent and defendant, as well as a constant. (For computational reasons, model 3 only estimates the student and student interaction terms but still conditions on, i.e., "controls for," the others.) The omitted, baseline categories are, respectively, judge, *Affirm*, and sympathetic defendant (i.e., the constant estimates the probability that a judge affirms the conviction of the sympathetic defendant under *Affirm*). Model 1 reports coefficient estimates, while models 2 and 3 report odds ratios. In each case, 95% confidence intervals are reported in brackets.

All three models show the same basic pattern consistent with table 2. The critical estimates are those of the interaction terms "student × *reverse*" and "student × unsympathetic" and their joint Wald or score test in the bottom row, which are highlighted by bold face. The interaction terms estimate the differences in precedent and defendant effect sizes, respectively, between judges and students. Relative to judges, students are estimated to move in the direction indicated by precedent 47 percentage points (model 1) or with almost 20:1 odds (models 2 and 3) more often than judges. Students are also estimated to be much less influenced by the defendant: the student interaction term estimate almost exactly offsets the baseline defendant estimate (which estimates the effect for judges). The 95% confidence intervals exclude equality with judge effects for precedent in all three models and for defendant in model 1. More to the point, the joint hypothesis that neither precedent nor sympathy effects differ between judges and students is soundly rejected in all three models at *p*≤0.007, the best estimate probably being *p*=0.002 from the exact model 3. In short, the estimated experimental effects for judges and students in our experiment do not merely happen to fall on different sides of significance thresholds but are substantively and statistically significantly different.

Table 3: Regressions

	Dependent Variable: Affirmed		
	(1) OLS (coefficients)	(2) Logit (odds ratios)	(3) Exact Logistic (odds ratios)
Precedent: *reverse*	0.15 [-0.14,0.44]	2.85 [0.41,19.93]	(conditioned on)
Defendant: unsympathetic	0.35 [0.06,0.64]	10.15 [1.03,99.64]	(conditioned on)
Student	0.41 [0.14,0.68]	12.1 [2.28,64.0]	10.9 [1.70,83.7]
Student × *reverse*	**-0.47** **[-0.80,-0.13]**	**0.05** **[0.00,0.48]**	**0.06** **[0.00,0.77]**
Student × unsympathetic	**-0.34** **[-0.67,-0.00]**	**0.10** **[0.01,1.28]**	**0.13** **[0.00,1.80]**
Constant	0.50 [0.28,0.73]	0.89 [0.27,2.95]	(conditioned on)
N	122	122	122
Joint *p*-value for student interaction terms (× *reverse* & × unsympathetic)	**0.002**	**0.007**	**0.002**

95% confidence intervals in square brackets. The joint *p*-value for the interaction terms is from a Wald test (models 1 and 2) or score test (model 3).

Yet another way to think about the difference between judges and students is to ask how likely one would draw a sample of 31 students from the student population that would give results as or more extreme than the actual sample of 31 judges. This approach quantifies the concern that we simply drew a highly unusual sample of judges. How unusual would the judge sample have to be if the populations of judges and students actually decided identically? We cannot answer this question precisely without access to the population of students, but we can approximate the answer with our sample of students. Specifically, we can *estimate* the student *population* affirmance probabilities for each treatment combination by the corresponding affirmance proportions in our student *sample*. Using these estimates, we derive the probability distribution of all possible contingency tables with 31 students distributed across the 2×2 treatment combinations like the judges in the actual judge sample. In other words, we analytically derive the full bootstrap distribution of a 31-student sub-sample. The probability of drawing a student (bootstrap) sample with an estimated effect in the same direction and a Fisher exact *p*-value as low as the judges' is 0.1% for sympathy and 0.02% for precedent. In short, it would be extremely unlikely to draw a sample as seemingly unmoved by precedent but moved by defendant sympathies as our judge sample.

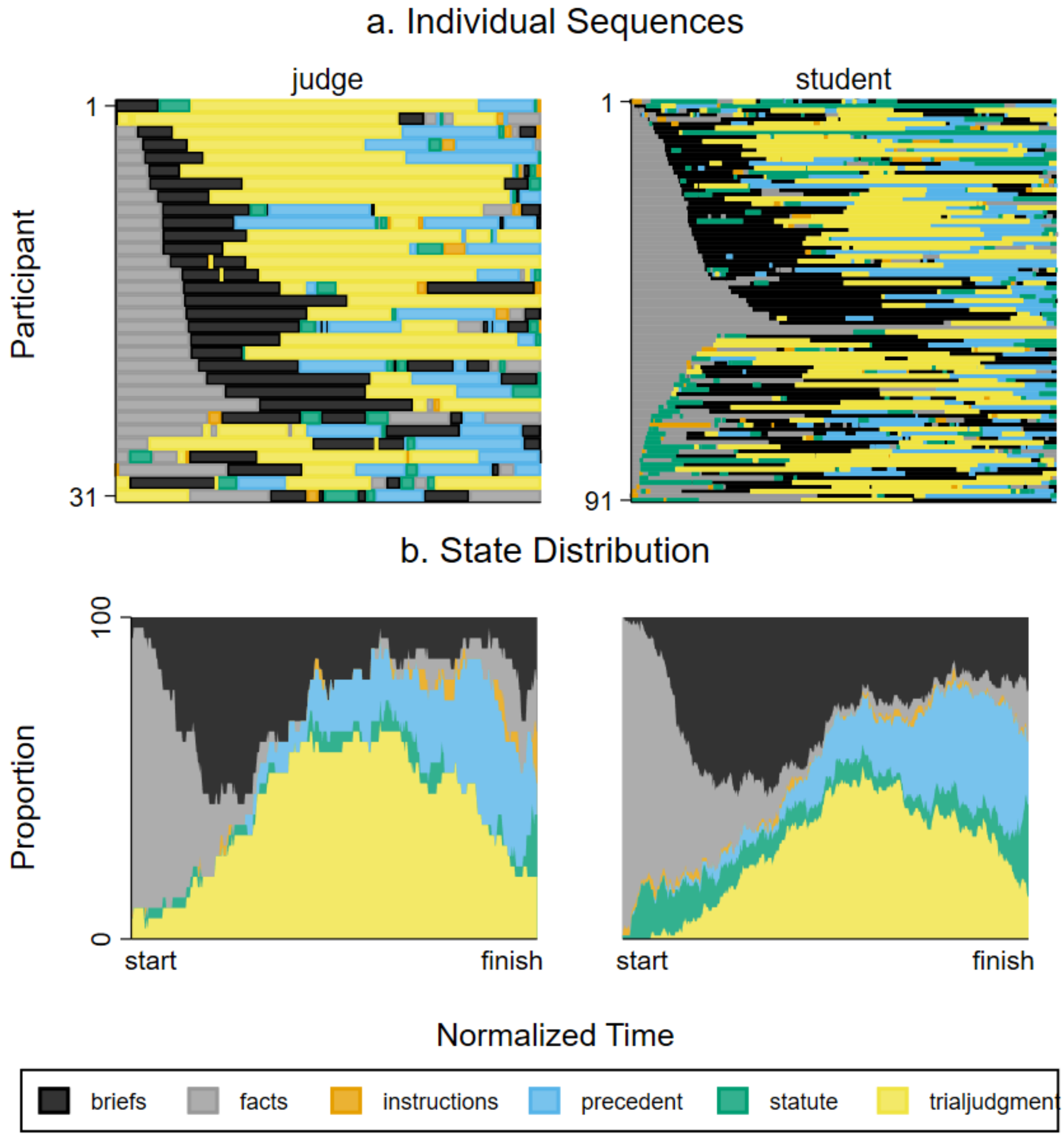
4.2 Process: Document View Paths

Given that judges and students differ strongly in their decisions, it stands to reason that they also differ in their reasoning process. That process is not observable as such, but we have access to a proxy:

participants’ document view paths, as recorded by our system in 10-second increments as participants worked on the case. We introduced this method in Spamann et al. (2021) and refer to that paper and its supplementary materials for a more thorough discussion.

Figure 1 plots the view paths for judges on the left and students on the right. The horizontal axis is time, which is normalized by each participant’s total time, so that all paths are of the same length. Panel 1.a plots individual view paths for each participant, where participants are aligned on the vertical axis. Panel 1.b plots the state distribution, i.e., what proportion of participants of the respective group has a particular document open at a given time—here vertical height corresponds to that proportion. The “resolution” we show here is the document that participants were working on at any given moment. We recorded the paths at the paragraph level but have found this too detailed and heterogeneous to make sense of.

Fig. 1: Document View Paths by Participant Type



By participant type, the graph shows (a) individual sequences (stacked vertically) and (b) the state distribution of documents live on participants' screen from start to finish (horizontal axis).

The document view paths give some indication that judges and students do not think alike. Panel 1.b makes it easy to see that judges tended to spend a larger proportion of their time with the trial judgment than students (40% vs. 27%), and that students consulted the statute much more frequently, especially early on. That said, the wide variety of different individual sequences shown in each group in panel 1.a suggests that individual differences dwarf differences between groups.

To evaluate the differences between judges' and students' document paths rigorously taking into account the ordering of document views, we employ the method of Spamann et al. (2021). The basic idea is to check if the paths within each group (judges or students, as the case may be) are more similar to each other than to paths in the other group. We measure pairwise similarity by the Levenshtein edit distance between two paths, discretized to 500 time periods.[24] Following Studer et al. (2011), we then compare the pseudo-R^2 of the actual groups to randomly labelled groups of equal size.[25] In 100,000 random labellings, we only once observed a pseudo-R^2 as high as that of the actual groups, i.e., we reject equality of judges' and students' distributions of document view paths at p=10^{-5}.

Given the novelty of the method, it is difficult to gauge substantive as opposed to statistical significance. On the one hand, the pseudo-R^2 is only 0.016, consistent with the visual impression that individual differences dwarf group differences. Specifically, a pseudo-R^2 of 0.016 means, roughly, that the distance between the two groups is only 1.6% of the average distance between individual participants. On the other hand, this pseudo-R^2 is almost twice as high as that between common and civil law jurisdictions and almost two fifths of that between individual countries considered in Spamann et al. (2021). When we nearest-neighbor-match students to judges based on the Levenshtein distance of their document view paths, students no longer differ from judges with respect to defendant effects but still differ strongly by precedent effect. This hints that the differences in sequences are meaningful, but the small effective sample of only 24 students (one student could be the closest neighbor to multiple judges) does not allow strong inferences.

4.3 Written Reasons

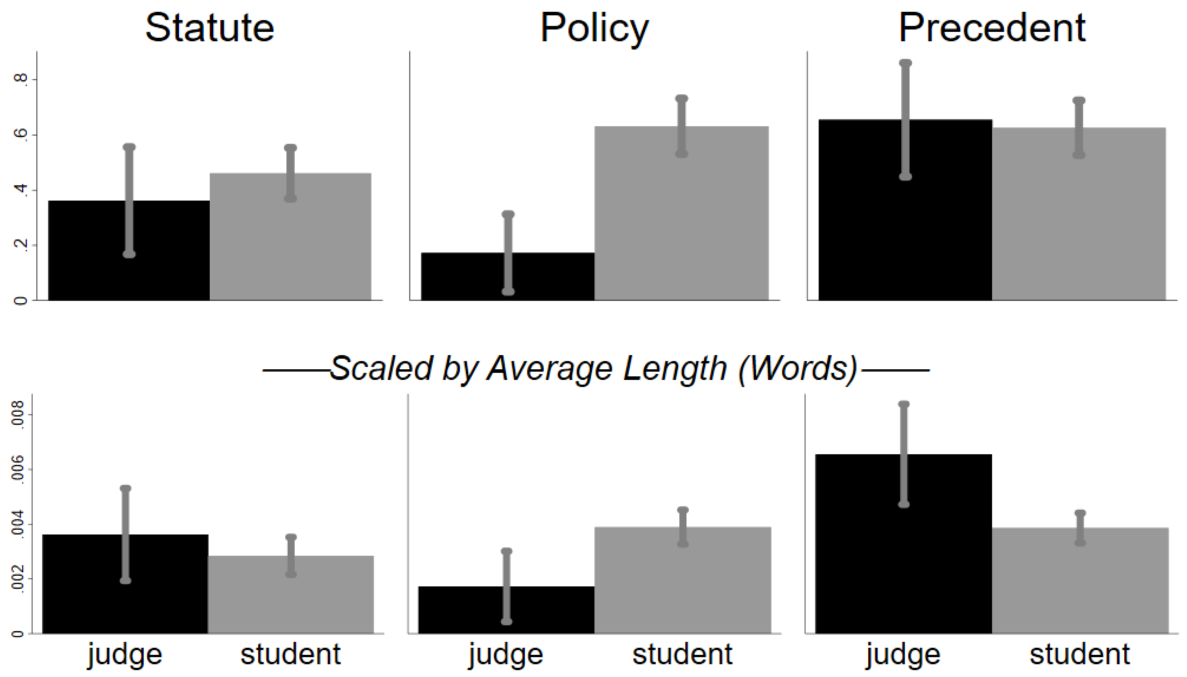
Finally, we have written reasons from all participants except two judges. Caution is required in comparing them because judges had to write with pen and paper or on the iPad's touchscreen keyboard, whereas students could type on whatever device they chose to use, presumably a personal computer with a physical keyboard. It is thus not surprising that the average student wrote considerably more words (162) than the average judge (100).

Two research assistants independently coded the arguments in participants' judgment reasons using the coding protocol in Appendix A.4. The protocol asked for binary answers to the questions "did the participant mention" (1) "precedent, even without the name of the precedent;" (2) "the statute"; (3) "policy**,** i.e. what impact the judgment might have on future behavior?" The two coders' answers were in almost perfect agreement. For the 122 observations, the coders disagreed only 2/3/1 time(s) for precedent/statute/policy, respectively. Where the coders disagreed, we retained the mean (0.5) for purposes of figure 2 and statistical tests (but we get qualitatively identical test results if dropping these observations or treating them as a third category).

[24] We obtain virtually identical results with a modified edit distance (substitution cost 1.5 instead of 1) and/or when eliminating two participations of less than five minutes (who are students; they likely did not take the experiment seriously).

[25] We gratefully acknowledge usage of Halpin's (2017) implementation.

Fig. 2: Prevalence of Reasons by Type



Means and 95% bootstrap confidence intervals of mentions of specified reasons by individual participants. In the upper panel, the mean is taken over indicators whether an individual participant mentioned the feature in the reasons. In the lower panel, the type mean is divided by the average number of words in reasons by that type.

Figure 2 shows the prevalence of the three arguments by participant type (judge or student) with 95% bootstrap confidence intervals. The top panel shows the raw prevalence, i.e., the proportion of the respective participant type that used the argument. As mentioned, judges on average wrote shorter reasons and hence might have had less opportunity to mention some argument. To address this, the bottom panel divides the prevalence by the average number of words written by that participant type.

A multivariate test strongly rejects equality of the joint distribution of the three arguments between judges and students (MANOVA using Pillai's trace, $p \leq 10^{-4}$ regardless of scaling). The primary driver is mentions of policy, which were considerably more frequent by students than judges (63% vs. 17%, Fisher exact $p<0.001$), a difference that remains substantively and statistically significant even after scaling. By contrast, students and judges barely differed in referring to the statute, scaled or not. For precedent, there is no difference in unscaled prevalence (about 63% of each type mentioned it), and judges even have a higher scaled prevalence ($t=2.83$, $p=0.008$). This is ironic because only the students' decisions differed by precedent (supra 4.1).

5 Conclusion

The behavior of law students and federal judges differs significantly in our study, both statistically and substantively. Observationally, we find major differences in document view paths and written reasons. More to the point, we obtain diametrically opposed experimental treatment effects in the two groups: judges' decisions differ by the bias treatment factor but not the precedent treatment factor, while students' decisions exhibit the opposite pattern. Like any experimental finding, ours could be due to chance, but we estimate the likelihood of that to be one in 500 or less (*supra* 4.1). The upshot is unfortunate for the experimental study of judicial decision-making: law students, who are easy to recruit, cannot stand in as experimental subjects for judges, who are very difficult to recruit.

This finding leads to two related research question, one substantive and one methodological. The substantive question is when and how law students morph into judges in the course of their career, and/or what distinguishes the law students that ultimately become judges. That students matched on document sequence judge more similarly but still differently from judges might suggest that both selection and maturation are involved but this is at best a hint, among other things due to the small effective sample size for this exercise (*supra* 4.2). The methodological question is whether other actors in the legal system, particularly practicing attorneys and arbitrators, behave like judges in experiments (as in Kahan et al. 2016), such that they could stand in for them (arbitrators being, however, similarly hard to recruit). It is also possible, albeit unlikely, that the federal judges in our experiment differ from state judges, or at least certain kinds of state judges (e.g., elected judges). These remain fascinating questions for future study – but they themselves will need to surmount the difficulty of recruiting judges.

References

Braman, Eileen, and Thomas E. Nelson. 2007. Mechanism of Motivated Reasoning? Analogical Perception in Discrimination Disputes. American Journal of Political Science 51:940-956.

Engel, Christoph, and Nina Grgić-Hlača. 2021. Machine Advice With a Warning About Machine Limitations: Experimentally Testing the Solution Mandated by the Wisconsin Supreme Court. Journal of Legal Analysis 13:284-340.

Ericsson, K. Anders. 2018. An Introduction to the Second Edition of The Cambridge Handbook of Expertise and Expert Performance: Its Development, Organization, and Content. In K. Anders Ericsson et al. eds., The Cambridge Handbook of Expertise and Expert Performance, 2nd ed., 3–20. Cambridge: Cambridge University Press.

Feess, Eberhard, and Roee Sarel. 2018 Judicial Effort and the Appeals System: Theory and Experiment. Journal of Legal Studies 47:269-294.

Furgeson, Joshua R., Linda Babcock, and Peter M. Shane. 2008a. Do a Law's Policy Implications Affect Beliefs About Its Constitutionality? An Experimental Test. Law and Human Behavior 32:219-227.

Furgeson, Joshua R., Linda Babcock, and Peter M. Shane. 2008b. Behind the Mask of Method: Political Orientation and Constitutional Interpretive Preferences. Law and Human Behavior 32:502-510.

Gilbert, Michael D. 2011. Does Law Matter? Theory and Evidence from Single-Subject Adjudication. Journal of Legal Studies 40:333-365.

Guthrie, Chris, Jeffrey J. Rachlinski, and Andrew J. Wistrich. 2007. Blinking on the Bench. Cornell Law Review 93:1-43.

Halpin, Brendan. 2017. SADI: Sequence analysis tools for Stata. The Stata Journal 17:546-572.

Holste, Lukas, and Holger Spamann. Forthcoming. Experimental Investigations of Judicial Decision-Making. In Kevin Tobia ed., The Cambridge Handbook of Experimental Jurisprudence. Cambridge: Cambridge University Press.

Iuliano, Jason, and Avery Stewart. 2016. The New Diversity Crisis in the Federal Judiciary. Tennessee Law Review 84:247-299.

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

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

Klerman, Daniel, and Holger Spamann. 2024. Law Matters – Less Than We Thought. Journal of Law, Economics & Organization 40:xx-yy.

Liu, John Zhuang, Lars Klöhn, and Holger Spamann. 2021. Precedent and Chinese Judges: An Experiment. American Journal of Comparative Law 69:93-135.

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

Mehrotra, Devan V., Ivan S. F. Chan, and Roger L. Berger. 2003. A Cautionary Note on Exact Unconditional Inference for a Difference between Two Independent Binomial Proportions. Biometrics 59:441-450.

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

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

Redding, Richard E., and N. Dickon Reppucci. 1999. Effects of Lawyers' Socio-Political Attitudes on Their Judgments of Social Science in Legal Decision Making. Law and Human Behavior 23:31–54.

Schauer, Frederick. 2010. Is there a Psychology of Judging? In David E. Klein and Gregory Mitchell eds., The Psychology of Judicial Decision Making 103-116. Oxford: Oxford University Press.

Spamann, Holger, and Lars Klöhn. 2016. Justice is Less Blind, and Less Legalistic, Than We Thought: Evidence from an Experiment with Real Judges. Journal of Legal Studies 45:255-280.

Spamann, Holger, Lars Klöhn, Christophe Jamin, Vikramaditya Khanna, John Zhuang Liu, Pavan Mamidi, Alexander Morell, and Ivan Reidel. 2021. Judges in the Lab: No Precedent Effects, No Common/Civil Law Differences. Journal of Legal Analysis 13:110-126.

Studer, Matthias, Gilbert Ritschard, Alexis Gabadinho, and Nicolas S. Müller. 2011. Discrepancy Analysis of State Sequences. Sociological Methods & Research 40:471-510.

Wistrich, Andrew J., Jeffrey J. Rachlinski, and Chris Guthrie. 2014. Heart Versus Head: Do Judges Follow the Law or Follow Their Feelings? Texas Law Review 93:855-923.

Appendix: Student-Specific Study Materials

Please refer to the online appendix of Spamann & Klöhn (2016) for the remaining study materials and judge-specific study materials.

A.1. Student Informed Consent

Differences between student rounds 1 (left box) and 2 (right box) are shown in boxes.

Participation is voluntary You are eligible to participate in this study if you are a 2L or 3L at [school 1], at least 18 years of age, and have not taken either of international criminal law or international humanitarian law.	**Study Eligibility** You are eligible to participate in this study if you are a 2L or 3L at a US law school, you are at least 18 years of age, and you have not taken either international criminal law or international humanitarian law. **Participation is voluntary**

It is your choice whether or not to participate in this research. If you choose to participate, you may change your mind and leave the study at any time. Refusal to participate or stopping your participation will involve no penalty or loss of benefits to which you are otherwise entitled.

What is the purpose of this research?

The goal of the study is to learn about the process of legal reasoning and the role of various legal materials therein. I am not testing your knowledge of, or opinions about, particular legal issues.

In the future, I plan to run the same study in other jurisdictions and compare the results.	For scientific reasons, I am not providing your more details about the study questions and design at this point. You will be fully debriefed at the end.

How long will I take part in this research?

Your participation will take approximately one hour to complete.

What can I expect if I take part in this research?

As a participant, you will be asked to judge a fictitious yet highly realistic international law case. I expect and indeed hope that you are completely unfamiliar with the applicable law. Relevant legal materials will be provided to you online. The site will log all of your activity on the site, i.e., which materials you consult when (but it will not record your IP address). At the end, you will also be asked to sketch the reasons for your judgment in a paragraph.

	In addition, there are a few exit questions including gender and class year.

What are the risks and possible discomforts?

	There are not expected to be greater than minimal risks from participation in this research.

If you choose to participate, the effects should be comparable to those you would experience from viewing a computer monitor for an hour and, if you are not using a tablet, using a mouse or keyboard.

Are there any benefits from being in this research study?

At …	There are no direct benefits expected to you as a consequence of participation in this study except that, at …

… the end of the study, I will provide a thorough explanation of the study and of our hypotheses. If you wish, you can send an email message to hspamann@law.harvard.edu, and I will send you a copy of any manuscripts based on the research (or summaries of our results).

Will I be compensated for participating in this research?

You will receive a $20 Amazon gift code if you complete the study and provide your name and your [round 1: school 1 / round 2: official law school] email address (the name is required for tax purposes …

… and the email to send you the gift code).	… , and the email must be provided to Amazon to send you the gift code).

If I take part in this research, how will my privacy be protected? What happens to the information you collect?

I will not collect identifiable information, including IP addresses, except your name and email if you claim your compensation. If you do give your name and email to claim your compensation, it will be kept confidential and entirely separate from the results of the study, and it will be available only for tax auditors; no identifiers will link the names and emails to the study data.	You may take this study at a time and place of your choosing. I will not collect identifiable information, including IP addresses, except your name and email if you claim your compensation. If you do give your name and email to claim your compensation, (1) no key will link the names and emails to the study data, (2) the names and emails will be stored on my password protected and encrypted hard-drive and cloud backup and shared only with tax auditors and, solely for purposes of sending the gift code, Amazon.

If I have any questions, concerns or complaints about this research study, who can I talk to?

The researcher for this study is Assistant Professor Holger Spamann who can be reached at (617) 496-6710, hspamann@law.harvard.edu, or Harvard Law School, 1525 Massachusetts Avenue, Cambridge MA 02138. Contact him

- If you have questions, concerns, or complaints,
- If you would like to talk to the research team,
- If you think the research has harmed you, or
- If you wish to withdraw from the study.

This research has been reviewed by the Committee on the Use of Human Subjects in Research at Harvard University. They can be reached at 617-496-2847, 1414 Massachusetts Avenue, Second Floor, Cambridge, MA 02138, or cuhs@fas.harvard.edu for any of the following:

- If your questions, concerns, or complaints are not being answered by the research team,
- If you cannot reach the research team,
- If you want to talk to someone besides the research team, or
- If you have questions about your rights as a research participant.

	Click here to print a copy of this information for your records. [print page hyperlink] **Click here to indicate your agreement to participate and proceed to the study. [proceed hyperlink]** If you do not want to proceed, you can exit the experiment by closing this browser window.

A.2: Instructions

{differences between judges and students, and between rounds of students, in curly braces}

Please imagine you are an appeals judge in the case Prosecutor v. [NAME] pending at the International Criminal Tribunal for the Former Yugoslavia (ICTY). This case is fictitious but very closely resembles an actual case recently decided by the ICTY. The ICTY is an international tribunal with the power to prosecute persons responsible for serious violations of international humanitarian law committed in the territory of the former Yugoslavia since 1991 in accordance with the provisions of the ICTY Statute. [ONLY HALF OF PARTICIPANTS WILL SEE: As an international tribunal, the procedure of the ICTY combines elements from common law and from civil law systems, some of which may seem unfamiliar to you.]

You have already presided over several hearings. The prosecution and the defence have now submitted their final appeals briefs and agreed on a list of agreed facts.

Your task is to judge whether the defendant is or is not guilty of aiding and abetting various war crimes by the [RELEVANT MILITARY GROUP] on the territory of Bosnia-Herzegovina in the years 1992-1994.

In reaching your judgment, you will be able to peruse the aforementioned briefs and the list of agreed facts. I recommend you read these in full. The briefs link to other documents, namely the decision of the trial court below, a recent decision by the Appeals Chamber in another case, and the statute establishing the ICTY. These other documents are very long. You will not have time to read them in full, but you may pursue a handful of further passages that you deem particularly relevant.

Please do NOT access any information on another device such as your smart phone, and please do NOT talk to {Judges: your neighbors until the study is completed} {Students: anyone about this case until you have completed it}.

You have 50 minutes to reach a decision and submit a brief summary of your reasoning {Judges: , either on this computer or on a separate piece of paper marked with your participant number, which will be randomly generated at the end of the study}. To help you keep track of time, a clock on the screen will count down the 50 minutes.

By clicking on the button below, you will proceed to an index page with all the documents provided. You can at any time return to this introduction or to the index page by clicking the relevant link at the top of the page. {students round 2: For technological reasons, however, you will not be able to open more than one browser window at once.}

A.3 Student Exit Questionnaire

{differences between student rounds 1 and 2 are marked in curly braces}

Please answer the following {7/6} short questions:

1. What proportion of your colleagues do you think decided the case as you did?
2. [FOR THOSE WHO ACQUITTED: Assume you were outvoted and the appeals chamber upheld the conviction.] What sentence would you find appropriate?

   For comparison, Charles Taylor, the former president of Liberia, was sentenced to 50 year in prison by an international tribunal for aiding and abetting widespread brutality in Sierra Leone that included murder, rape, the use of child soldiers, the mutilation of thousands of civilians and the mining of diamonds to pay for guns and ammunition. On the other hand, Razim Delic, the chief of staff of the Army of Bosnia and Herzegovina during the war, was sentenced to only three years by the ICTY for his failure to prevent members of his army from committing crimes against captured civilians and enemy combatants (murder, rape, torture).

   {round 1: [form field was prepopulated with 10, 25, or 40]}

   {round 2: To avoid misunderstandings, please write the time units (years or months) explicitly. For example, if you thought that [10/25/40] years were an appropriate penalty, you should write either “[10/25/40] years” or “[120/300/480] months.” [numbers of years and corresponding months randomized]}
3. Did you have any previous knowledge of international criminal law? Y/N
4. Did you recognize any of the names or places in the case? Y/N
5. {round 1: Who taught your criminal law class in your 1L year? / round 2: What is your class year – 2L or 3L?}
6. What is your gender? Female Male {round 2 only: rather not say]}
7. {round 1: Did you encounter any technical difficulties or inconveniences while navigating this experiment? If yes, could you please describe these difficulties?}

A.4 Coding Protocol for Judgment Reasons

<u>REFUSAL</u>: Did the participant explicitly refuse to submit judgment?

1. This includes any judge who *explicitly* notes that they did not want to give judgment but clicked through by mistake, to express their disapproval, etc.
2. <u>If not</u>: did the participant express *explicit* **reservations** about submitting a judgment (in particular because of the short time available)?

<u>ERROR</u>: Did the Participant make a Clerical Error in Entering Judgment, i.e., does the Judgment (0/1) Clearly not correspond to what the Participant Wanted, Given the Reasons?

<u>MISUNDERSTANDING</u>: Are there clear indications that the participant misunderstood the task (of judging the case on appeal)?

We are NOT judging the quality of their written work product but merely looking for indicators that they did not do what they were supposed to be doing in this study. In particular, an answer that is merely cryptic, inconclusive, or even inconsistent is not evidence of a misunderstanding. Bear in mind that the judges were working under time pressure, and time may have run out before they finished their sentence or before they could delete obsolete words they had written earlier. Also, judges may purposefully use conclusory language, distort the facts, overstep their role, etc. -- happens in real life all the time.

Similarly, participants need not couch their reasons explicitly in the language of reviewing the lower court's findings. In other words, it is not evidence of a misunderstanding that a participant did not explicitly frame their argument as a review of the lower court's findings. First, most appeals courts around the world review questions of law de novo, so discussing the legal evaluation of facts established by the trial court is consistent with the appeal's judge role whether or not it is framed as a "review" of the trial court. Second, some appeals courts in the world are allowed to review the entire case de novo, see next note. Third, we must again make allowance for the time pressure and the rhetorical short cuts that it may have inspired.

1. NB: Making factual findings is not per se evidence of a misunderstanding. First, the ICTY Appeals Chamber is technically allowed to review "error of fact" under Art. 25(1)(b) of the ICTY statute (provided the factual error "has occasioned a miscarriage of justice"). Second, many appeals courts in the (civil law) world are allowed to review facts de novo, and we purposefully did not specify a standard of review. Also see the next question.
2. If you answer yes to, or had doubts about, this question 2., please include a note explaining why

PRECEDENT – did the participant mention precedent, even without the name of the precedent

STATUTE – did the participant mention the statute

POLICY – did the participant mention policy**,** i.e. what impact the judgment might have on future behavior?