# LAW AND PSYCHOLOGY GROWS UP, GOES ONLINE, AND REPLICATES

Kristin Firth, David A. Hoffman & Tess Wilkinson-Ryan*

Abstract:

Over the last thirty years, legal scholars have increasingly deployed experimental studies, particularly hypothetical scenarios, to test intuitions about legal reasoning and behavior. That movement has accelerated in the last decade, facilitated in large part by cheap and convenient Internet participant recruiting platforms like Amazon Mechanical Turk. The widespread use of online subjects, a practice that dramatically lowers the barriers to entry for experimental research, has been controversial. At the same time, the field of experimental psychology is experiencing a public crisis of confidence widely discussed in terms of the "replication crisis." At present, law and psychology research is arguably in a new era, in which it is both an accepted feature of the legal landscape and also a target of fresh skepticism. The moment is ripe for taking stock.

In this paper, we bring an empirical approach to these problems. Using three canonical law and psychology findings, we document the challenges and the feasibility of reproducing results across platforms. We evaluate the extent to which we are able to reproduce the original findings with contemporary subject pools (Amazon Mechanical Turk, other national online platforms, and in-person labs). We partially replicate all three results, and show marked similarities in subject responses across platforms. In the context of the experiments here, we conclude that meaningful replication requires active intervention in order to keep the materials relevant and sensible. The second aim is to compare Amazon Mechanical Turk subjects to the original samples and to the replication samples. We find, consistent with the weight of recent evidence, that the Amazon Mechanical Turk samples are reasonably appropriate for these kinds of scenario studies. Subjects are highly similar to subjects on other online platforms an in-person samples, though they differ in their high level of attentiveness. Finally, we review the growing replication literature across disciplines, as well as our firsthand experience, to propose a set of standard practices for the publication of results in law and psychology.

* Kristin Firth is a graduate student at the University of Chicago's Booth School of Business. David Hoffman is a Professor of Law at the University of Pennsylvania School of Law. Tess Wilkinson-Ryan is a Professor of Law and Psychology at the University of Pennsylvania School of Law.

## INTRODUCTION

In the last decade, the costs of doing law and psychology research have fallen dramatically. Unsurprisingly, the supply of experimental papers—especially hypothetical vignette studies—is up, both in peer-reviewed journals and in student-run law reviews.[1] Scenario studies now inform a wide range of subjects, including international law,[2] torts,[3] criminal procedure,[4] contracts,[5] and securities.[6] It has never been easier to run, and publish, a quickie experiment.[7]

Commentators have identified two concerns about this state of the scholarship. First, they worry that contemporary law and psychology studies are too dependent on a narrow, unusual subject pool: Amazon Mechanical Turk ("MTurk"). MTurk is extraordinarily inexpensive—some researchers pay as little as ten cents per subject—and does not require any sort of institutional affiliation. Many scholars have conjectured that the MTurk subject pool is unrepresentative of the broader population in both observable and unobservable ways.[8] MTurk participants are clearly younger and more liberal than the American population. They are also, by definition, willing to do online survey research for very small payments, an odd enough trait that it raises questions about other hidden idiosyncrasies. Ease of use also raises suspicions of false positives—if it's very easy to test a hypothesis repeatedly, the odds of at least one statistically significant result are, of course, higher. Skepticism about MTurk is now widespread among consumers and arbiters of empirical legal studies, from student editors to peer referees.

---

[1] *See* Adam J. Berinsky, Gregory A. Huber & Gabriel S. Lenz, *Evaluating Online Labor Markets for Experimental Research: Amazon. com's Mechanical Turk*, 20 POL. ANAL. 351, 351 (2012) ("Interest in experimental research has increased substantially in political science."); *See generally* Shari Seidman Diamond & Pam Mueller, *Empirical Legal Scholarship in Law Reviews*, 6 ANN. REV. L. SOC. SCI. 581, 589 (2010) (discussing the rise of experimental research in law).

[2] *See, e.g.,* Adam Chilton & Dustin Tingley, *Why the Study of International Law Needs Experiments*, 52 COLUM. J. TRANSNAT'L L. 173 (2013) (discussing how experimental research methods are increasingly utilized in international law).

[3] *See generally* Jennifer K. Robbennolt & Valerie P. Hans, The Psychology of Tort Law (2016).

[4] *See, e.g.*, Avani Mehta Sood, Cognitive Cleansing: Experimental Psychology and the Exclusionary Rule, 103 GEO. L.J. 1543 (2015) (using scenario study to conduct experiments on exclusionary rule); *See also* Holger Spamann & Lars Klöhn, *Justice is Less Blind, and Less Legalistic, Than We Thought: Evidence from an Experiment with Real Judges*, 45 J. LEGAL STUD. 255 (2016) (investigating judicial decision making through scenario simulation).

[5] *See, e.g.,* George S. Geis, *An Experiment in the Optimal Precision of Contract Default Rules*, 80 TUL. L. REV. 1109 (2006) (conducting experimental study to determine optimal precision of contract default rules); *See also* George S. Geis, *Empirically Assessing Hadley v. Baxendale*, 32 FLA. ST. U. L. REV. 897 (2005) (using empirical assessment to analyze key case in contract law); Yuval Feldman & Doron Teichman, *Are all Contractual Obligations Created Equal?*, 100 GEO. L.J. 5 (2011) (using experimental surveys to study responses to a breach of contract).

[6] *See, e.g.,* Jill Fisch & Tess Wilkinson-Ryan, *Why Do Investors Make Costly Mistakes?: An Experiment on Mutual Fund Choice*, 162 U. Pa. L. Rev. 605 (2014).

[7] Others have made similar points. *See* Kathryn Zeiler, *The Future of Empirical Legal Scholarship: Where might we go from Here?*, 66 J. LEGAL EDUC. 78, 84-87 (2016) (discussing lack of interest in methods of introspection within legal academy).

[8] *See infra* at text accompanying notes 34 through 35.

The second source of skepticism of law and psychology arises from new debates about the value of psychological research more generally. The field of psychology is in the midst of a mid-life crisis of sorts, stemming from recent claims that a surprising fraction of experimental psychological findings are not replicable—or at least that they are not reliably replicable in their original form. The crowd-sourced "Replication Project" and its offshoots represent a profound challenge to many sub-disciplines inside of psychology, as they seem to call into question foundational findings –not to mention the careers they've launched.[9] Indeed, though psychology has come under particularly public scrutiny, the failure to reproduce scientific results is under active discussion across the academy, in economics, in medicine, and even in physics.

In this paper, we bring together these two sources of skepticism about law and psychology, taking seriously both the critiques and the justifications of the field and its methods. The paper offers a case study of a replication project, as well as a set of systematic comparisons of subject characteristics across survey platforms. Specifically, we replicate three canonical law and psychology papers, using MTurk, the commercial survey firm Survey Software International ("SSI"), Prolific Academic, a new online service designed for researchers, and an in-person lab run by a university. Using thousands of respondents, including in-person and nationally representative pools, and with different scenarios, we draw some lessons for experimental law and psychology.

Researchers in the social sciences have offered ample evidence that MTurk subjects respond to classic psychology experiments in predictable and reliable ways—at least as compared to other common sample populations. We find the same here. We also find that they are significantly more attentive than subjects in other subject pools.

In our view, allaying some of these more arithmetic concerns about MTurk opens up the deeper conversation about experimental psychology in legal studies. Resistance to MTurk and like online platforms arises in part from a worry that as the costs of doing research falls, the quality will suffer. That reasonable fear ought to be addressed on its own terms. How should consumers of this field distinguish between good work and bad, signal and noise? We do not think the answer is to sort based on subject pool, but agree that the demands of quality control are higher in a world of universal access to randomizing survey software and cheap online subjects.

This paper also makes a more theoretical contribution, drawing on our research experience . We argue that context, specifically temporal context, is an unappreciated treatment for experiments about law. It is a truism that the present moment influences

[9] *See infra* at text accompanying notes 10 through 18.

peoples' views about legal institutions and rules—and yet it is also easily overlooked. Social, political, and legal change over time make rote replication of law and psychology experiments not just challenging but almost irrelevant. Though we did in fact reproduce many findings of the canonical papers we studied by using their original materials, we came to view that approach as counterproductively rigid, and even unfair to the original authors. "Replication" of law and psychology demands that researchers focus on the underlying mechanism of action and work to translate those insights to a modern vernacular.

We conclude with some proposals for best practices in the field. Unlike others, we do not recommend pre-commitment to particular research designs. We do, however, recommend transparency in piloting, public data storage, multiple scenario testing, and diverse platform subject series. These modest recommendations would go a long way to assuaging concerns about MTurk, and would shore up the foundations of law and psychology as it continues its explosive growth.

The paper proceeds in four Parts. We begin in Part I by describing the nature of the problem facing the field. Part II presents our method and describes the survey populations we sample. Part III reports our results, and Part IV offers a synthesis.

## I. REPRODUCIBILITY AND GENERALIZABILITY IN LAW AND PSYCHOLOGY

This part brings together two cross-currents that have recently created turbulence for law and psychology: the "replication crisis" in its home discipline of psychology, and the nagging doubts about MTurk, the most popular modern source of subjects for legal experiments.

### *A. The Replication Crisis in Psychology*

The replication problem transcends scientific fields. In 2005, John Ioannidis published a bombshell essay entitled, "Why Most Published Research Findings are False."[10] In the simplest terms, his damning critique of the hypothesis-to-publication pipeline is that when researchers are in search of a particular result with a p-value of less than .05, someone will find that result by chance alone one in twenty times—and that's the trial that will get published. Ioannidis's essay was largely about biomedical science—a particularly vivid narrative emerged from the study of genomics, which "learned the importance of replication the hard way"[11]—but his critiques and others like it have had a

[10] *See* John P. A. Ionnidis, *Why Most Published Research Findings are False,* 2 PLOS MED. 696 (2005).

[11] *See, e.g.,* Peter Kraft, Eleftheria Zeggini, and John Ioannidis, *Replication in Genome-Wide Association Studies*, 24 STAT. SCI. 561 (2010).

serious impact in the social sciences, including psychology.

In 2015, *Science* published "Estimating the Reproducibility of Psychological Science," the report of a crowd-sourced replication project led by Brian Nosek.[12] Each of ninety-eight teams of researchers, drawn from many academic institutions, attempted to replicate a different previously-reported result in psychology. The Open Science Collaborative chose 100 recent papers from three highly-regarded psychology journals. In most cases, researchers were instructed to attempt to replicate the last reported study in the chosen paper. Each study was replicated once. Replication teams used the original effects sizes to determine the sample size required for adequate power to detect the effect. They obtained original materials where possible, established the protocol for the replication, and contacted original authors for feedback on the replication method's fidelity to the original study. They ran the study, reported the results, and judged whether or not the original result had replicated via five converging measures, from statistical significance to subjective evaluation. The outcome was startling and heavily publicized; less than half of the studies were replicated. Indeed, only about 35% of reported effects were significant at the $p < .05$ level in the replication. The headline in *Nature* was "Over Half of Psychology Studies Fail Reproducibility Test." *The Atlantic* covered it in an article entitled, "Psychology's Replication Crisis Can't be Wished Away." *Slate*, which has covered the replication debate extensively, described the Nosek findings in an article called "Everything is Crumbling."[13]

The Open Science Collaborative project clearly captured the public imagination, to the chagrin of many in the field, who argued that the studies were underpowered, poorly conceived, and overhyped. In a follow-up paper Nosek and his colleagues wrote that their study provides no grounds for drawing "pessimistic conclusions about reproducibility."[14] Skeptics of their methods and their conclusions replied in an open response that "We hope they will work as hard to correct the widespread public misperceptions of the article as they did on the article itself."[15]

The OSC project, and subsequent debates and critiques, crystallized two serious

---

[12] *See* Brian Nosek et al., *Estimating the Reproducibility of Psychological Science*, 349 SCIENCE 3451 (2015).

[13] *See* Daniel Engber, *Everything is Crumbling,* SLATE March 2016, available at http://www.slate.com/articles/health_and_science/cover_story/2016/03/ego_depletion_an_influential_theory_in_psychology_may_have_just_been_debunked.html

[14] Brian Nosek et al., *Response to Comment on "Estimating the reproducibility of psychological science,"* 351 SCIENCE 1037, 1037 (2016).

[15] Daniel T. Gilbert, Gary King, Stephen Pettigrew & Timothy D. Wilson, *A Response to the Reply to Our Technical Comment on "Estimating the Reproducibility of Psychological Science,"* PsychCentral (March 2016), https://psychcentral.com/blog/wp-content/uploads/2016/03/gkpw_response_to_osc_rebutal.pdf.

points of theoretical and empirical contention inherent to replication: fidelity and power. How close to the original design is close enough? And what is the threshold level of evidence required for declaring a failure to replicate?

### 1. Power and Effect Size

The power issue—the question of how much evidence of the predicted effect is required to replicate or to fail to replicate—drives much of the debate in the replication literature. The question is how many observations adequately power a study. Conceptually, power matters more in replication than it does in new research, because replication studies want to be able to say either that a finding did replicate, or that it did not. The latter claim is essentially a claim about a null result. Failed replication attempts yield the conclusion that a particular finding does not exist; failed original research yields no more than a claim that the researchers are unable to rule out the null hypothesis. To draw the bolder conclusion, the research team must be able to make strong claims that the study design would find the result if the phenomenon were real.

There is an attractive arithmetic solution to this problem. Rather than ask if the phenomenon is real, replication teams can choose to define the phenomenon of interest by both direction (does this cause people to move up or down on this measure?) and magnitude (up a lot or a little?; down a lot or a little?). The OSC project used the original effect sizes of the chosen studies to determine how to power the replications. They identified a sample size big enough to capture with 80% certainty an effect if the magnitude of true effect—say, the difference between the treatment and the control—was anywhere in the 95% confidence interval of the original effect size. In a rebuttal to the Open Science paper, Daniel Gilbert and colleagues compared its approach to that of the Many Labs project.[16] The Many Labs project took 16 published studies and attempted to replicate each study in 36 different labs.[17] The Many Labs project replicated 14 of its 16 chosen studies, in part because of the dramatic increase in power.

The argument against the effect-size approach of the OSC is that it misses true differences that are robust to replication but smaller than originally estimated. Because published results are more likely to overestimate an effect size than to underestimate it—since underestimated effect sizes are less likely to get published—replications that depend on original effect size will systematically be underpowered. This dispute about effect sizes is fundamental. Psychologists are increasingly explicit about the limited role

---

[16] Daniel Gilbert, G. King, S. Pettrigrew & Timothy Wilson, *Comment on "Estimating the Reproducibility of Psychological Science,"* 351 SCIENCE 6277 (2016).

[17] Richard Klein et al., *Investigating Variation in Replicability: A "Many Labs" Project*, 45 SOCIAL PSYCHOLOGY 152 (2014).

that effect size can and should play in replication. Uri Simonsohn has articulated the problem as follows: "Psychological theories are almost exclusively qualitative rather than quantitative—predicting sign rather than magnitude of effects—and hence are not well equipped to help us identify when an effect is too small to be of theoretical interest." When we describe our approach in the methods section of the paper, we will explain our adoption of Simonsohn's "small telescopes" approach to powering replication.[18]

**2. Fidelity**

The second challenge to replication effort is the extent to which the replication is reproducing the original study in a meaningful way. Is the replication faithful enough to the original research that it should be expected to capture a result if one exists? One of the most salient critiques of the original Open Science Collaboration report was a short list of particularly egregious infidelities, along with the observation that even under the authors' own criteria for successful replication, the replication protocols endorsed by the original authors had a significantly higher chance of producing similar results than the unendorsed (low fidelity) protocols. The solution to the fidelity problem appears straightforward at first—if researchers are trying to reproduce a finding, they should reproduce the stimuli as near as feasible to the original. Like the power question, though, the fidelity question is more difficult than it appears on its face. In some cases, literal fidelity will be counterproductive. An easy illustration is studies with monetary values involved. Replicating a study with values that fail to account for inflation will change the meaning of the scenario and the response variables. A more complex problem might arise from a study with outdated facts in a hypothetical. Should the replication team attempt to update with a similar set of facts? How should they assess similarity? As we discuss in Part IV, these questions are particularly trenchant in the context of the psychological analysis of law and legal behavior, where political and social context are inherent to the phenomena of interest.

*B. Amazon Mechanical Turk in Experimental Legal Studies*

In part because of the relative youth of the discipline, legal scholars have not generally focused on the replicability of their experimental studies. Rather, methodological concerns are centered around subject recruitment, and in particular, the expanding use of the MTurk platform. MTurk is an increasingly popular way that legal scholars have recruited subjects.[19] We see four main drivers of this popularity:

---

[18] Uri Simonsohn, *Small Telescopes: Detectability and the Evaluation of Replication Results*, 26 PSYCH. SCIENCE 559 (2015).

[19] We found over 175 experimental papers in the JLR database on WL, 2/3 in the last two years, using MTurk as a subject recruitment device.

- **Cost**: Per subject fees on MTurk range from a few cents to a few dollars, turning on the nature of the task and the researchers' motivations and ethical constraints. By contrast, costs for other convenience samples (like Qualtrics) can be in the $5.00 range, and the costs for curated nationally representative samples are in the $10-$20 per complete range.
- **Convenience**: Legal scholars rarely (if ever) have easy access to the undergraduate psychology majors who make up the backbone of traditional psychology experimental work. MTurk requires no collaboration with psychology professors, is available at a moment's notice, and at the click of the button, and is relatively easy to use.
- **Diversity**: MTurk subjects are more diverse than college student samples, across various demographic variables.
- **Standardization:** As MTurk becomes more popular, it is easier to explain, support, and defend its use by reference to other published research.

But is MTurk a useful source—that is, "useful" for producing high-quality social science—of survey respondents? Two papers from the beginning of this decade were the prime motivators in MTurk's adoption by survey researchers. The Paolacci, Chandler, and Ipeirotis ("PCI") paper on MTurk, published in JUDGMENT AND DECISIONMAKING, is the urtext for social psychology researchers using MTurk.[20] PCI studied MTurk at an important point in its history: the population of workers (who we will interchangeably call "subjects") had recently globalized due to Amazon's shifting payment practices, leaving researchers unsure about the sample's demographics and performance.[21] PCI concluded—in a line which may have launched a thousand research projects—"[i]n sum, U.S. workers on MTurk are arguably closer to the U.S. population as a whole than subjects recruited from traditional university subject pools."

After examining the subject pool, PCI detailed several advantages to running experiments on MTurk, and then reported the results of a comparative test of MTurk workers responses to three classic JDM heuristic-elicitation tasks with responses gathered from the internet more generally, and with college subjects. All three groups produced roughly equivalent responses, though MTurkers were more responsive (attentive) than the general internet pool.[22] Thus, at least with respect to classic heuristic tests, MTurk purportedly could offer cheaper, faster, and more reliable data than existing populations.

---

[20] With over 2500 google scholar citations.

[21] Gabriele Paolucci, Jesse Chandler and Panagiotis Ipeirotis, *Running Experiments on Amazon Mechanical Turk*, 5 JDM 411, 412 (2010).

[22] PCI, *supra* note 21, at 417.

The second major paper, published in 2011 by Buhrmester, Kwang, and Gosling, had the inviting title, *Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?*[23] BKG replicated some of the findings from PCI's paper. They compared answers to personality questionnaires from 3000+ subject sample from MTurk with those from a large alternative internet sample.[24] Finding that the MTurk sample produced psychometrically valid personality scales, the authors concluded that "[m]Turk participants are at least as diverse and more representative of non-college populations than those of typical Internet and traditional samples. Most important, we found that the quality of data provided by MTurk met or exceeded the psychometric standards associated with published research."[25] BKG's paper has been cited over 3,000 times.

There are many subsequent papers on the representativeness and utility of the MTurk worker population.[26] Canvassing these papers, it is hard to fairly generalize about the exact MTurk sample at any given time, since not only does the sample change over time, but Amazon estimates that more than 500,000 individuals across the world have registered as workers.[27] Of that pool, recent work has suggested that no more than 10,000 (US-centered) workers may be active at any one time.[28] These "superturkers" are significantly more active—and sophisticated —than the ordinary workers.[29]

The gender split identified by PCI has balanced out in recent years.[30] Respecting other demographic characteristics, results vary. One study, for example, concludes that workers are "wealthier, younger, more educated, less racially diverse, and more Democratic than national samples" and less religiously affiliated.[31] Others are more equivocal, concluding that apart from workers' relative youth and liberalism, they closely approximate a nationally representative sample.[32]

---

[23] Michael Buhrmester, Tracy Kwang and Samuel D. Gosling, *Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?* 6. Persp. Psych. Sci. 3 (2011). BKG's paper has been cited over 3000 times in google scholar.

[24] *Id.* at 4.

[25] *Id.* at 5.

[26] An important synthesis can be found in KIM SHEEHAN AND MATTHEW PITTMAN, THE ACADEMIC'S GUIDE TO USING AMAZON'S MECHANICAL TURK: THE HIT HANDBOOK FOR SOCIAL SCIENCE RESEARCh 13 (Melvin & Leigh, 2016).

[27] Sheehan and Pittman, *id.* at 14.

[28] *See* Neil Stewart et al., *The Average Laboratory Samples a Population of 7,300 Amazon Mechanical Turk Workers*, 10 JDM 479, 480 (2015).

[29] One study suggests that Workers follow a power law—10% completing 80% of the HITS. Pam Mueller, Jesse Chandler and Gabrielle Paolacci, *Advanced uses of Mechanical Turk in Psychological Research,* Presentation to the Society for Personality and Social Psychology - January 28, 2012, available at https://experimentalturk.files.wordpress.com/2012/01/mueller-spsp-2012.pdf

[30] Sheehan and Pittman, supra note 26, at 17.

[31] Andrew R. Lewis et al., *The Non(religion) of Mechanical Turk Workers*, 54 J. SCI. STUDY OF RELIG. 419, 420 (2015).

[32] Berinsky et al., *Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk*, 20 Pol. Sci.

The crucial question is whether differences in the subject pool, whether or not observable, lead MTurk workers to distinctive patterns of behavior and thought. This validity concern motivated Professor Dan Kahan's popular series of posts against MTurk samples on the Cultural Cognition Blog.[33] Kahan provocatively began his challenge by noting that while MTurk's price is purportedly a mark in its favor, "thought experiments are even cheaper. But they are *not valid*."[34] Several categories of Kahan's criticism recur in the literature:

*First*, critics argue that there may be unobservable oddities in Turkers' political ideology. Citing the literature reviewed above, Kahan suggests that not only is it a problem that there are fewer conservatives in the MTurk sample than the national population, but that the conservatives that remain are not typical of American conservatives (i.e., less partisan, less religious, etc.)[35] Research since the publication of Kahan's critique, however, found that MTurk conservatives "share the same personality traits and values as conservatives drawn from high-quality national samples."[36]

*Second*, Kahan argued that MTurk samples are non-naïve, particularly among the population of superturkers. The key paper, by Chandler, Mueller and Paolacci, found that a small number of workers had taken over twenty studies a week and more than 3,000 overall. These superturkers see the same scales repeatedly[37], and sometimes discuss the experiments they are undertaking.[38] While it is fairly easy to screen against workers who have taken a particular researcher's experiment, it is much more difficult to

---

351 (2012); Connor Huff and Dustin Tingley, *Who are these People: Evaluating the Demographic Characteristics and Political Preferences of Amazon Mechanical Turk Participants,* Research and Politics September 2015 1-12, available at http://rap.sagepub.com/content/2/3/2053168015604648.full.pdf+htmlA2012; Clifford, S., Jewell, R. M., & Waggoner, P. D. (2015). Are samples drawn from Mechanical Turk valid for research on political ideology?. *Research & Politics*, *2*(4), 2053168015622072.

[33] In the interests of full disclosure, one of us is a member of the Cultural Cognition Project, a co-author of Kahan, and a very, very occasional contributor to that blog.

[34] Dan Kahan, *A Pigovian tax solution (for now) for review/publication of studies that use M Turk samples,* CULTURAL COGNITION PROJECT (Jun. 9, 2015, 8:02 AM), http://www.culturalcognition.net/blog/2015/6/9/a-pigovian-tax-solution-for-now-for-reviewpublication-of-stu.html.

[35] See Dan Kahan, *Fooled twice, shame on who? Problems with Mechanical Turk Study Samples, Part 2,* CULTURAL COGNITION PROJECT (Jul. 10, 2013, 9:30 AM), http://www.culturalcognition.net/blog/2013/7/10/fooled-twice-shame-on-who-problems-with-mechanical-turk-stud.html.

[36] Scott Clifford, Ryan M. Jewell and Phillip D. Waggoner, *Are Samples Drawn from Mechanical Turk Valid for Research on Political Ideology,* Research and Politics 2015 (1500 subject sample findings that MTurk conservatives "share the same personality traits and values and conservatives drawn from high-quality national samples" while "MTurk liberals appear to hold more characteristically liberal values and political attitudes.") available at http://rap.sagepub.com/content/2/4/2053168015622072.full.pdf+html

[37] Jesse Chandler, Pam Mueller & Paolacci, *Nonnaivete Among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers*, 46 BAH. RES. METHOD. 112 (2014).

[38] This can lead to the reduction of effect sizes. Jesse Chandler, Gabriele Paolacci, Eyal Peer, Pam Mueller & Kate A. Ratliff, *Using Nonnaive Participants Can Reduce Effect Sizes*, 26 PSYCHOL. SCI. 1131 (2015).

effectively screen against superturkers. Researchers may be right to worry that subjects less familiar with surveys and vignette research in general might respond in systematically different ways than those who are quasi-professional questionnaire-takers.

*Third,* MTurk workers act differently from other samples in that they are *highly* motivated to have their work accepted. In our experience as researchers who use MTurk, we've been struck by how seriously subjects take HIT rejection, and how vigorously they will argue to have work accepted that did not comply with the relevant experimental instructions, regardless of the low stakes involved. It's no surprise, therefore, that as compared to college undergraduates, Turkers tend to pay more attention to survey prompts.[39] In the experiments described in this paper, as the reader will see, the differences in attentiveness were dramatic. This is both good (less noise) and bad (less normal).[40] Also, because Turkers care about reputation, experimental manipulations targeting subject motivation may be difficult to pull off.[41]

*Fourth,* there is a worry that Turk workers identifying as US citizens may not, in fact, be located in the US. One study found that 6% of purportedly US-based workers in fact hailed from Eastern Europe and India.[42] While, to a degree, such deceit can be controlled through IP-address monitoring, those techniques may not be perfect.[43]

The upshot of these arguments is that MTurk workers will produce different results on important psychological measures than subjects recruited from representative samples. Evidence for this proposition is, to date, quite limited. Krupnikov & Levin compared the performance of college undergrads, MTurk workers, and diversely recruited national pool on four classic political science attitude surveys.[44] They found that the MTurk workers produced answers that differed from both the college pool and the nationally representative pool, both in terms of political attitude's magnitude and, sometimes, direction. However, more recent work has found that controlling for

---

[39] Eyal Peer, Joachim Vosgerau & Alessandro Acquisti, Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk, 46 BEHAV. RES. 1023 (2014) (high-reputation AMT workers rarely failed attention checks and provided high-quality data).

[40] Adam Berinsky et al., Separating the Shirkers from the Workers? Making Sure Survey Respondents Pay Attention on Self-Administered Surveys, 58 AM. J. POL. SCI. 739 (2014).

[41] A recent paper suggests a decrease in reliability of MTurk samples especially when controls against bots and other low-quality responses were not imposed. Steven V. Rouse, *A Reliability Analysis of Mechanical Turk Data*, 43 COMPUTERS IN HUM. BEHAV. 305 (2015) (recommending anti-bot devices).

[42] Danielle N. Shapiro, Jesse Chandler & Pam A. Mueller, *Using Mechanical Turk to Study Clinical Populations*, 1 CLINICAL PSYCHOL. SCI. 213 (2013)

[43] Kahan, *supra* note 36.

[44] Yanna Krupnikov & Adam Seth Levin, *Cross-Sample Comparisons and External Validity*, J1 J. EXP. POL. SCI. 59 (2014).

observable individual differences, MTurk respondents do *not* appear to differ fundamentally from population-based respondents in unmeasurable ways.[45] Bartneck and his co-authors, for example, find trivially small differences between Turk populations and a campus sample in a task rating emotional expressions.[46] And there are no papers that find systematic differences in classic social psychology main effect findings between MTurk samples and those from either college undergraduates or nationally representative and curated pools.

The latest word in this series of papers, by Eyal Peer and coauthors,[47] compares MTurk respondents with other online platforms: CrowdFlower, a corporate-oriented marketing crowdsourcing platform, and Prolific Academic, a platform founded by graduate students and intended for research, along with a traditional offline participant pool. The underlying tasks were adopted "from prominent studies in psychology" (i.e., the Asian Disease framing task) and included attention checks and demographic questions. Overall, as compared to MTurk, CrowdFlower respondents were faster, but paid less attention and were less reliable. Prolific and MTurk produced results essentially similar to the offline samples, but at a fraction of the cost and in less time.

As this brief synopsis suggests, there is actually a great deal of evidence that MTurk workers are not unobservably unrepresentative, and that experiments using MTurk are not unusually susceptible to external validity concerns. We would still be skeptical that MTurk subjects are reliable proxies for individuals in the exercise of their professional judgment—judges, corporate directors, lawyers.[48] And we think care ought to be exercised in using MTurk subjects for the study of individual differences, especially where we have reason to suspect heterogeneous treatment effects based on ideology or partisanship. But the weight of the evidence seems to be in favor of MTurk as a reasonable subject pool; at this point, the burden of persuasion is on those who are arguing otherwise.

---

[45] Kevin E. Levay, Jeremy Freese, and James N. Druckman, *The Demographic and Political Composition of Mechanical Turk Samples*, Sage Open Repository 2015, http://sgo.sagepub.com/content/6/1/2158244016636433.full-text.pdf+html; Kevin J. Mullinix, Thomas J. Leeper, James N. Druckman and Jeremy Freese, *The Generalizability of Survey Experiments*, 2 J. EXP. POL. SCI. 109 (2016) (finding that ATE in convenience and nationally representative samples are similar); Alexander Coppock, *Generalizing From Survey Experiments Conducted on Amazon Mechanical Turk: A Replication Approach*, Pol Sci. Res. Meth. (forthcoming), available at http://alexandercoppock.com/papers/Coppock_generalizability.pdf (same).

[46] Christoph Bartneck et al., *Comparing the Similarity of Responses Received from Studies in Amazon's Mechanical Turk to Studies Conducted Online and with Direct Recruitment*, PLOS ONE 10:e0121595 (2015)

[47] Eyal Peer, Laura Brandimarte, Sonam Samat & Alessandro Acquisti, *Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research*, 70 J. EXP. SOC. PSYCH. 153 (2017).

[48] The idea that expert decision making is reliably different than non-experts decisions is associated with Howard Margolis. *See* HOWARD MARGOLIS, DEALING WITH RISK: WHY THE PUBLIC AND THE EXPERTS DISAGREE ON ENVIRONMENTAL ISSUES 35 (1996)

* * *

Against this overall backdrop—greater than ever production of law and psychology research, along with new and newly trenchant criticisms of psychological methods—we decided to begin for law and psychology the project that is ongoing in psychology and political science: try to replicate some studies on MTurk and other samples, and compare the results across time and platform. We are explicitly *not* making a general claim about the reproducibility of experimental legal studies as a field (nor could we, with three studies). Rather, we would like to show, through replication of canonical legal experiments on a variety of survey platforms, the potential and the pitfalls of replication in law and psychology. We also intend to test the prediction that MTurk subjects behave similarly to subjects recruited via more costly platforms.

## II. METHOD

### *A. Which Experiments?*

We chose to concentrate on replicating studies from three early, well-cited experimental papers which were published prior to the origination of MTurk. We considered and piloted multiple studies and settled on three.[49] The first study, from Jennifer Robbennolt, involves an accident and the effect of apologies on a victim's willingness to settle a claim.[50] The second study, from Jeffrey Rachlinski, investigates the framing effects of litigant roles—plaintiffs (gain frame) and defendants (loss frame).[51] The third, from Cass Sunstein and co-authors, involves differential assessments of physical and financial harms depending on the implicit categorization of the harm.[52] Each of these

[49] We ultimately chose three studies that fit our criteria of being 1) well-known/well-cited and 2) reasonably within our shared areas of expertise. Our original goal was to choose 3 to 5 studies in total. We began with four studies to pilot using MTurk subjects. Those four studies were the three reported in detail here, as well as Janice Nadler's "Flouting the Law" (2005). We ultimately decided we could not move forward with the Nadler replication, because the materials were highly specific to a particular political moment—and the social/political meanings of the stimuli was central to the hypotheses being tested. The underlying phenomena being described in that paper is complex enough that it seemed clear that dated materials would undermine replication attempts, and any updating would require sustained thought and attention from researchers with a deep understanding of the relevant issues. We were guided by the question, "Would a failed replication be informative as to whether or not the posited relationship exists?" We determined that we could not answer that question in the affirmative without a much more serious investment for a study (extensive piloting to test for relevant political views, identification of divisive issues with cognizable stakes for our subjects, etc.) that was about social and political values.

[50] Jennifer Robbennolt, *Apologies and Legal Settlement: An Empirical Examination*, 102 MICH. L. REV. 460 (2003), 484-490 (describing the "Effects of Apologies on Settlement Decisionmaking" study)

[51] Jeffrey J. Rachlinski, *Gains, Losses, and the Psychology of Litigation*, 70 S. CAL. L. REV. 113 (1996), 135-140 (describing Study One methods and results).

[52] Cass Sunstein, Daniel R. Kahneman, David Schkade and Ilana Ritov, *Predictably Incoherent Judgments,* 54 STAN. L. REV. 1153 (2002).

is well-known in the field, has important implications for law, and was sufficiently methodologically transparent to permit replication.[53]

*B. How Many Subjects?*

> "It is generally very difficult to prove that something does not exist; it is considerably easier to show that a tool is inadequate for studying that something. With a small-telescopes approach, instead of arriving at the conclusion that a theoretically interesting effect does not seem to exist, we arrive at the conclusion that the original evidence suggesting a theoretically interesting effect exists does not seem to be adequate."[54]

We used the power calculation from Simonsohn's influential 2014 paper, "Small Telescopes: Detectability and the Evaluation of Replication Results."[55] The paper makes the argument that power calculations for replication of behavioral studies ought to give us a number of observations for which a null result tells us something meaningful. The "small telescopes" framework is motivated by an intuitively compelling metaphor: if a large telescope fails to detect a phenomenon reported by a scientist with a much smaller telescope, that failure is meaningful—things observable with small telescopes ought to be observable with large telescopes. This method essentially combines an estimation of effect sizes, as in traditional power analysis, with basic hypothesis testing in which the hypothesis being tested is that the true effect is undetectably small, or zero. "For the latter objective, accepting the null hypothesis, we ask the following question, 'If the true effect size is 0, how many observations do we need to have an 80% chance of concluding that the effect is undetectably small?'" The answer is surprisingly simple; a replication *with about 2.5 times as many observations as the original study* has about 80% power to reject the hypothesis that the effect is undetectable. This approach is the most conceptually coherent that we have identified, and we adopted it here.

---

[53] In choosing these studies, we chose three studies that we could run almost verbatim. And, indeed, for each study, we ran a pilot using the verbatim text, and also used the verbatim text with the SSI subjects. For the MTurk and in-person runs, we made (what we view as) minor updates in order to preserve or adapt the narrative. In each study, names of protagonists were either changed to sound more plausible/contemporary (a child named Joan was renamed Julia, for example). The gender-ambiguous "Pat" was changed to "Will." An "answering machine" became "voicemail." Some language was simplified to be accessible to all subjects. We view this as a mechanical replication, in the sense that the changes, which were overall very few, were made only where the language of the original seemed likely to confuse or inadvertently surprise contemporary online subjects.

[54] Simohnson, *supra* note 18.

[55] *Id.*

### C. *Which Platforms?*

The original studies were drawn from distinct samples and each used different recruitment and sampling methods.[56] We chose our replication platforms—MTurk, a commercial survey firm, and an in-person lab run by a university—to maximize our ability to compare common contemporary samples, as opposed to matching our sampling and recruitment to original studies (aware, of course, that this choice reduces fidelity to the original studies).

We ran each study using Qualtrics survey software, and recruited participants through three different subject pools: MTurk, SSI, and an in-person sample recruited by a behavioral lab at a midwestern university. On each platform we imposed (what we thought was) a pre-treatment, forgiving, attention check.[57] Surprisingly, while MTurk subjects failed that check at a 0.5% rate; in-person subjects failed at a 25.7% rate, and SSI subjects at a shocking 51.7% rate. We paid only those subjects who paid attention. For MTurk, subjects only completed one of the studies. Using the qualification feature of MTurk, we prevent prior study takers from viewing or completing new HITs. For SSI and the lab subjects, due to time and cost maximization goals, the same subjects completed the studies (in random order) in one session.

**MTurk:** Overall, we surveyed 1606 attentive MTurk subjects, over a one week period,[58] in September, 2016. The cost for these subjects was $0.70 to $1.40 per subject, for

---

[56] Robbenolt used a randomized email recruitment protocol to recruit university employees to complete a web-based questionnaire study. Rachlinski recruited undergraduate and law student to participate in an in-person lab study. Sunstein et al. used a commercial survey firm drawing on a pool of jury-eligible Texans.

[57] After an initial consent page, subjects saw a page with the following text:

**"This study seeks to understand how people process the questions that are being asked to them. There are many aspects of a person's behavior that are related to the way they answer questions.**

**One aspect is their ability to stay engaged throughout a survey and a person's willingness to read the directions fully. To make sure you are currently paying attention, for the question below we would like you to answer: "None of the above" Which of the following adjectives would you use to describe yourself?"**

After this text was a set of twelve attributes as well as the answer "none of the above" presented in randomized order. The way we were able to handle failed attention checks varied by platform.

On MTurk, if the subject failed the attention check we asked them to please return the HIT without completing it. We had already captured the participant's self-entered MTurk ID by this point, and would be able to see if they attempted to restart the survey. On SSI, if the subject failed the attention check they were redirected to a failure screen, were not paid for completing the task, and SSI provided a replacement participant. At the in-person lab, if subjects failed the attention check, they were given the opportunity to re-start the survey by the lab assistant. We were not initially aware that this was customary protocol, but decided to allow the procedure to continue as it usually progressed in this lab for the fairest comparison of that sample.

[58] For the MTurk platform, we wanted to run the studies at various times during the day and throughout the week; one concern people have expressed is that MTurk subjects may look very different during a weekday versus a weekend, because perhaps those with full-time jobs only answer surveys on weekends or at night. We randomly assigned each of the three studies to one weekday daytime slot, one weekday evening slot, and one weekend slot.

Each slot consisted of a day and an hour to start running the study, when we would initiate the HIT's availability on

a total of $3.50 for all three studies, inclusive of fees. Responding to concerns about naïveté, we attempted to remove anybody who may have seen more than one version of the survey.[59]

**Lab:** Our second platform was an in-person community lab run by a Midwestern university. There, we recruited 441 participants over a two-month period including a holiday break. Subjects came from a general community sample, including approximately 50% general public and 50% local college students in a metropolitan area. They were solicited to participate, offered $1 per five-minute block of time, and scheduled to go to the lab facility in person and take a subset of the studies being run at any particular time. The cost per participant for the in-person sample was $4.00.

**SSI:** Third, we recruited 665 subjects from SSI, which promises a nationally representative sample. SSI is a professional survey company that was founded in 1977 and has offered online samples for fifteen years. Participants were drawn from SSI's panels and various online communities, social networks, and websites of all types in the United States. [60] Functionally, SSI recruits subjects online (we believe using ads on various social network sites) by promising them "SSI points" in return for answering questions. One concern we had with the SSI sample – apart from the high number of recruited subjects who failed attention checks—is that it is not reliable for incentive-based experiments. SSI suggests that each point is roughly worth $0.01. But, in unpublished results of an experiment supporting previous work,[61] one of us found that while the average subject thought that 200 SSI points was worth $2.12, the standard deviation around that estimate $2.82. Quite simply, a fair number of attention-check complying SSI

---

MTurk. Start times were chosen assuming about 2-hours total to run the study. The choosing of slots was random, within a number of constraints. For the weekday daytime slots, we randomly chose three weekdays (Mon-Fri) and hours falling between 9am and 5pm for all time zones. For the weekday evening slots we excluded Friday evening, and randomly chose three other weekdays (Mon-Thur) and hours falling between 5pm and midnight for all time zones. For the weekend slots we randomly ordered one of each weekend days, and an extra randomly chosen weekend day (Sat or Sun), and randomly chose hours falling between 9am and 10pm for all time zones. Three of these 50 possible start times were blocked off due to unavoidable conflicts.

Appendix C describes demographic differences by day on the MTurk platform. No differences observed significantly interacted with the experimental manipulations in question.

[59] Of the 1662 participants who submitted a complete survey, 43 were removed because of having a duplicate ID or IP address and 13 were removed for not having a completed MTurk HIT. See Appendix B for full details of the process on how data are screened and removed.

[60] *See Consumer Online Survey Research*, SSI, https://www.surveysampling.com/solutions/data-collection/online-surveys/consumer/ (last visited Sept. 21, 2016); *see also* Netta Barak-Corren, *Does Antidiscrimination Law Influence Religious Behavior? An Empirical Examination*, 67 HASTINGS L.J. 957, 989 n.112 (2016) (using a nationally representative SSI sample in academic research).

[61] *Cf.* David A. Hoffman and Zev Eigen, *Contract Consideration and Behavior,* 85 GEO. WASH. L. REV. 351 (2017) (Hoffman and Eigen attempted to replicate the main findings of the incentive compatible paper using SSI, but found that subject confusion about their payment made such replication impossible).

subjects were not focusing on the survey prompts. The cost for the SSI sample was $5.64 per subject.

Demographic information across platforms is shown in Table 2.

**Table 2. Demographic information for each subject pool.**

| | **MTurk** | **SSI** | **Lab** | **2010 Census** |
|---|---|---|---|---|
| **N** | **1606** | **665** | **441** | |
| **Age:** Mean (range)<br>Median (25/75) | 37 (18-78)<br>34 (28/44) | 45 (18-90)<br>44 (32/59) | 35 (18-84)<br>31 (24/46) | 38.5 (median) |
| **Percent Male** | 48% | 48% | 62% | 49.2 |
| **Education:** | | | | |
| < high school - GED | 11% | 20% | 23% | 44% |
| some college | 39% | 34% | 48% | 19% |
| 4 year degree | 38% | 30% | 23% | 27% |
| grad/prof degree | 12% | 16% | 6% | 9% |
| **Household Income:** | | | | |
| < 50K | 52% | 40% | 80% | 43% |
| 50-100K | 38% | 39% | 17% | 29% |
| 100-150K | 8% | 13% | 2% | 14% |
| > 150K | 2% | 8% | 1% | 14% |
| **Employment:** | | | | |
| Full-time | 59% | 46% | 27% | N/A[62] |
| Part-time | 18% | 15% | 27% | |
| Retired | 4% | 19% | 4% | |
| Unemployed | 14% | 14% | 30% | |
| Full-time Student | 4% | 5% | 12% | |
| **Cost Per Subject (for three studies)** | $3.50 | $5.64 | $4.00 | |

## III. RESULTS

[62] Because the employment statistics from the 2010 census are not relevant to the composition of the 2016/17 labor pool, we omit them for clarity.

*A. Initial Results*

**1. Apologies**

This study opens with a scenario describing a bicycle accident.[63] Subjects read that while they were walking, a bicyclist ran into them and they ended up in the hospital. Subjects were randomly assigned to one of three types of apology conditions: *None*, *Partial*, or *Full*. In the *None* condition, subjects read that the bicyclist has not contacted them or apologized. In the *Partial* and *Full* conditions the bicyclist leaves a voicemail. In the Partial condition the voicemail contains a short message about being sorry. In the *Full* condition the voicemail contains a longer messaging, including the apology and an indication that it was "all" the bicyclist's fault. Subjects are first asked whether they would settle the case, then a series of other questions about the accident and apology.[64]

In this replication report, we are comparing primarily those in the *None* condition, who did not receive any apology, to those in the *Full* condition, who received the longer apology text.[65] We did not observe a significant main effect for willingness to settle, as measured by a straight comparison of means. With Turk subjects we found a significant increase in subjects who would "definitely accept" a settlement offer accompanied by an apology. All other underlying emotional/judgment response questions were significantly affected by the apology, in the predicted direction.

Put differently, as in the original study, the apology caused subjects to report that they would be more forgiving, less angry, and less condemning of the underlying conduct. However, we did not robustly replicate the main reported effect of increasing

---

[63] The full scenario text used on MTurk is provided in the Appendix. The text was slightly modified between platforms to improve readability.

[64] **Accept Settlement**: If Will were to agree to pay the amount of your medical bills that was not covered by insurance ($500) and to compensate you for your lost sick leave ($1,500), would you agree to settle the case?
**Sufficient Apology**: To what degree do you feel Will has offered a sufficient apology?
**Anger**: How angry would you be at Will?
**Bad Conduct**: How bad do you think Will's conduct was?
**Forgiveness**: How willing are you to forgive Will?
**Damage to Relationship**: How damaging will this incident be to your relationship with Will?
**Careful in Future**: How careful do you think Will will be when riding a bike in the future?
**Offer Makes Up for Injury**: To what degree would the settlement offered make up for your injuries?
These are all of the questions that we asked across all platforms and were in the original paper. Some platforms contained additional questions not reported here.

[65] Robbennolt mostly found that the *Partial* condition behaved similar to the *None* condition. We found that *Partial* sometimes behaved like *None* and sometimes behaved like *Full*. This result—that partial apologies may be interpreted differently depending on context, is consistent with other work, including Jennifer Robbenolt, *Apologies and Settlement Levers*, 3 J. Empirical Legal Stud. 333 (2006) (discussing divergent effects of a partial apology on subjects' intuitions). For simplicity of comparing main results, and reporting multiple platforms, we provide just the *None* and *Full* conditions here.

willingness to settle. The difference was in the predicted direction, and we found evidence to support the hypothesis (just comparing "definitely accepts" across condition), but overall the replication study did not elicit the same strong intuitions about settlement that were reported in the original.

Table 3 shows the significance tests for each variable, by platform. Table 4 shows mean differences across platforms. Figure 1 shows average treatment effects across platforms.

**Table 3. Robbennolt: Statistically significant differences between Full Apology and No Apology (all differences statistically significant in original)**

| | MTurk | SSI | Lab |
|---|---|---|---|
| N None/Full conditions | N=147/126 | N=97/99 | N=155/152 |
| Accept Settlement | Marginal, p=.1048 | No, p= .3494 | No, p=.5701 |
| Sufficient Apology | Yes, p<.0001 | Yes, p<.0001 | Yes, p<.0001 |
| Anger | Yes, p<.0001 | No, p= .5418 | Yes, p=.0040 |
| Bad Conduct | Yes, p<.0001 | No, p=.6537 | Yes, p=.0025 |
| Forgiveness | Yes, p<.0001 | Yes, p=.0536 | Yes, p=.0423 |
| Damage to Relationship | Yes, p<.0001 | Yes, p=.0103 | Yes, p=.0091 |
| Careful in Future | Yes, p=.0002 | Yes, p=.0012 | Yes, p<.0001 |
| Offer Makes Up for Injury | No, p=.1302 | No, p=.8634 | No, p=.3752 |

**Table 4. Robbennolt: Comparisons by Full Apology/No Apology treatment.**

| | Original | | MTurk | | SSI | | Lab | |
|---|---|---|---|---|---|---|---|---|
| Apology Condition | None | Full | None | Full | None | Full | None | Full |
| Accept Settlement | 52% | 73% | 64% | 68% | 65% | 68% | 60% | 56% |
| Sufficient Apology | 1.9 | 3.8 | 1.8 | 3.8 | 2.4 | 3.7 | 2.4 | 3.9 |
| Anger | 3.7 | 2.9 | 3.8 | 3.0 | 3.6 | 3.5 | 3.6 | 3.2 |
| Bad Conduct | 4.1 | 3.1 | 4.0 | 3.2 | 3.9 | 3.8 | 3.7 | 3.3 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Forgiveness | 3.6 | 4.2 | 3.4 | 4.0 | 3.6 | 3.9 | 3.7 | 4.0 |
| Damage to Relationship | 3.3 | 2.0 | 4.0 | 2.9 | 3.4 | 2.9 | 3.2 | 2.8 |
| Careful in Future | 3.5 | 4.2 | 3.7 | 4.1 | 3.7 | 4.2 | 3.9 | 4.4 |
| Offer Makes Up for Injury | 2.5 | 3.6 | 2.6 | 2.8 | 2.9 | 2.9 | 2.7 | 2.8 |

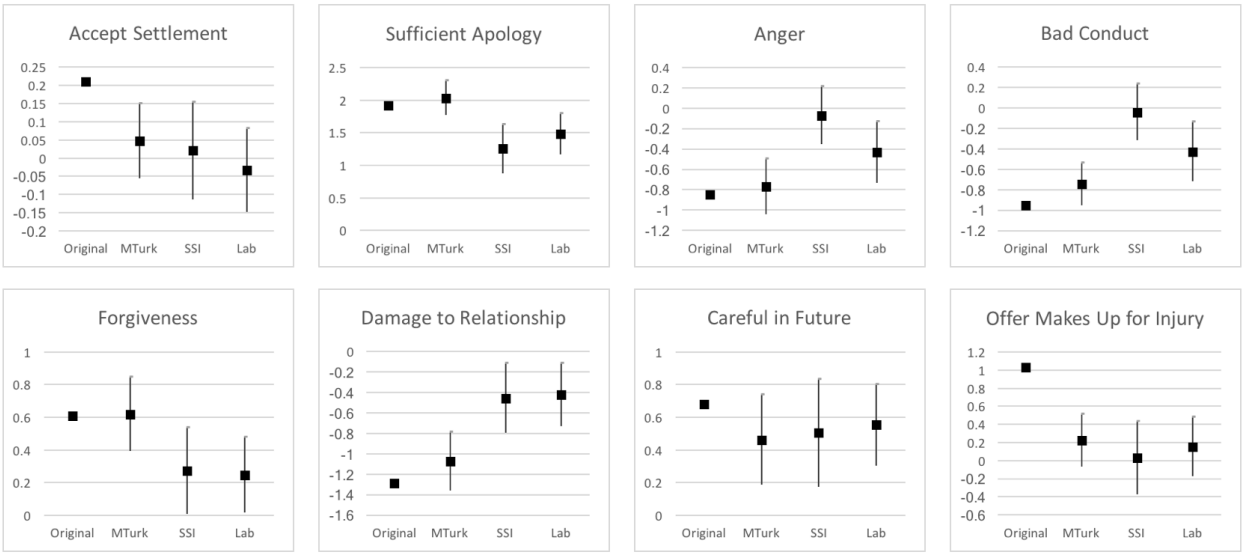


**Figure 1: Points are average treatment effects of the full apology (difference between full apology group means and no apology group means). Bars are the 95% confidence interval for the mean difference.**

### 2. Gain/Loss Frame

The Rachlinski study uses a scenario describing a property dispute.[66] Subjects are told to imagine they are in the role of an attorney representing one side of the dispute. They are randomly assigned to be told they represent the plaintiffs who have had their land encroached upon, or the defendants, who unintentionally encroached on the land. They are additionally randomly assigned to be given the impression that they are likely to win or lose (with 70% or 30% odds) at trial, by reading that a senior partner at their law firm is familiar with the judge who either is or is not an adamant supporter of property rights. Finally, they are given a settlement offer equal to roughly the expected value of the law suit, and are asked whether they would choose to settle. (Either by

[66] The full scenario text used on MTurk is provided in the Appendix. The text was modified slightly between platforms to improve readability.

receiving or paying the stated amount, depending on whether they represent the plaintiff side or defendant side.) The response choice is a binary "Yes" or "No" about whether they agree to the settlement.

In this study, subjects were compared across the same winning or losing condition. In other words, plaintiffs who were winning were compared against defendants who were winning. Meanwhile, plaintiffs who were losing were compared against defendants who were losing. Rachlinski's original paper found enormous differences between willingness to settle rates for defendants and plaintiffs—up to almost 50 percentage points. In our direct replication, we saw significant differences in the predicted direction among MTurk subjects, but the magnitude of the effect was diminished, hovering closer to 15 percentage points. Subjects in the community lab sample looked similar, but only one condition showed significant differences. The experimental manipulation had no discernable effect with subjects recruited from the commercial survey firm (though the small observed differences are in the predicted direction).

Table 5 shows the results of the direct replication for each platform, in both the Winning and Losing conditions. Both conditions yielded significant differences in the predicted direction in MTurk, neither replicated in SSI, and the Winning but not the Losing differences were significant at the community lab. Table 6 shows the values for comparisons across platforms.

**Table 5. Rachlinski: Did significant differences from original paper replicate? Statistically significant differences between Defendant and Plaintiff (all differences statistically significant in original)**

| | MTurk | SSI | Lab |
|---|---|---|---|
| N Losing (Ps / Ds) | 112 / 110 | 168 / 167 | 110 / 113 |
| N Winning (Ps / Ds) | 111 / 113 | 168 / 162 | 107 / 111 |
| Settling Losing Cases | Yes, p=.0114 | No, p=.4146 | No, p=.2947 |
| Settling Winning Cases | Yes, p=.0038 | No, p=.4003 | Yes, p=.0519 |

**Table 6. Rachlinski: Comparisons by Defendants and Plaintiffs**

| | Original | | MTurk | | SSI | | Lab | |
|---|---|---|---|---|---|---|---|---|
| Settling | % Ps | % Ds | % Ps | % Ds | % Ps | % Ds | % Ps | % Ds |
| Losing | 72.7 | 25.0 | 89.3 | 75.5 | 79.8 | 75.4 | 88.2 | 82.3 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Winning | 93.1 | 56.3 | 84.7 | 67.3 | 80.4 | 75.9 | 78.5 | 65.8 |

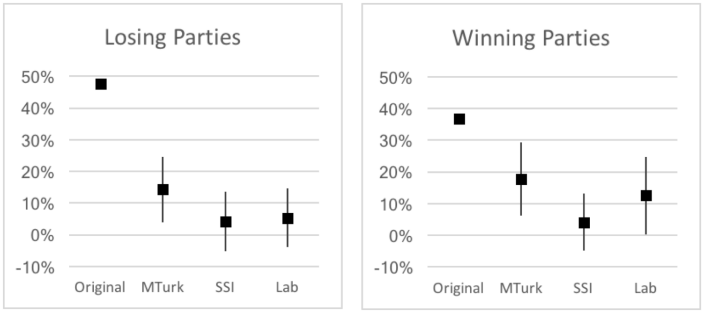


**Figure 2: Points are average treatment effects of the plaintiff framing (difference between plaintiff group means and defendant group means). Bars are the 95% confidence interval for the mean difference.**

### 3. Category Frame

The Sunstein et al. paper described subjects' reported valuations of a variety of harms. Subjects read about legally cognizable harms—"childhood safety cap fails; child needs hospital stay" or "distributor violates contract, damaging supplier's business"—as well as public "causes," like saving dolphins or ameliorating living conditions for migrant farmworkers. Overall, people were highly concerned about physical harms and harms to humans (as opposed to financial harms and harms to the environment). But the researchers hypothesized that financial and environmental harms would seem more serious when subjects were considering them in isolation—in the context of other kinds of financial or environmental harms, but not compared to other, more serious, categories of harm. For harms, they were asked how much money the wrongdoer ought to pay in punitive damages. For causes, they were asked how much they would themselves be willing to contribute to such a cause. The manipulation of interest was whether the items were presented in isolation, or in comparison across category. The authors hypothesized, and found, that physical and human harms were allocated more money relative to financial and environmental harms when the categories were presented together than they were in isolation.

Unlike subjects in the original study, subjects in the Sunstein replication were generally attentive to the more serious harm from the outset—they assigned a high value to the human/physical harms whether or not it was compared to other kinds of harms. For this reason, the effect of comparison was noticeably muted and largely non-significant, as measured by the interaction term of type of harm (personal vs. non-personal) and the evaluation mode (isolation vs. comparison). This is shown in Table 7.

**Table 7. Statistically significant interaction between Personal Injury or Financial Injury and evaluating first or second, ordinal regression.**

| | MTurk | SSI | Lab |
|---|---|---|---|
| N personal/financial first | N=362/358 | N=331/334 | N=220/221 |
| Legal cases (median) Item Type * Evaluation Mode | No, p=.7536 | No, p=.4336 | No, p=.6940 |
| Public cases (mean) Item Type * Evaluation Mode | Yes, p= .0510 | No, p=.290 | No, p=.316 |

However, the effect did replicate within-subjects—subjects who saw the personal/human harms at Time 2 were more likely to increase their allocations from Time 1 than subjects who saw the harms in the reverse order. Table 7b shows this breakdown—people were more likely to give more in the comparison condition (Time 2) if they were evaluating a human rather than an environmental or financial harm. This supplemental analysis is helpful in a study like this, because it suggests that the failure to replicate the original result has more to do with our subjects having different baseline intuitions rather than different responses to the experimental manipulations.

**Table 7b. Statistically significant differences between whether choosing to give more at time-2, based on injury type.**[67]

| | MTurk | SSI | Lab |
|---|---|---|---|
| T2 Personal > T1 Financial | 73% | 66% | 60% |
| T2 Financial > T1 Personal | 23% | 14% | 30% |
| More frequent if T2 Personal | Yes, p<.0001 | Yes, p<.0001 | Yes, p<.0001 |
| T2 Human > T1 Environment | 41% | 36% | 45% |
| T2 Environment > T1 Human | 23% | 31% | 26% |
| More frequent if T2 Human | Yes, p<.0001 | No, p=.1875 | Yes, p<.0001 |

[67] The original paper does not quantify within-subjects differences except to say, "Note also that when cases are compared across categories of harm, the evaluation of the more prominent item rises sharply, while that of the less prominent item declines slightly or stays about the same. The asymmetric effects of the comparison on more and on less prominent harms reflects a cognitive phenomenon that is well understood, but not central to our story." (1176-1177).

Another way to look at this is to consider the differences between the Time 2 estimates and Time 1 estimates, between subjects. A positive difference indicates the participant increased their damages for the second case, when engaging in joint evaluation. A negative difference indicates they decreased their damages for the second case. Participants who saw a physical or human harm second chose to increase their valuation.

**Table 7c. Differences in amount between time-1 and time-2, based on injury type.**

| | MTurk | SSI | Lab |
|---|---|---|---|
| T2-T1, T2=Personal | 400,000 | 400,000 | 100,000 |
| T2-T1, T2=Financial | - 100,000 | -497,500 | -2 |
| Greater if T2 Personal | Yes, p<.0001 | Yes, p<.0001 | Yes, p<.0001 |
| T2-T1, T2=Humans | 18.84 | 13.37 | 21.27 |
| T2-T1, T2=Environment | 6.161 | 10.94 | -10.77 |
| Greater if T2 Humans | Marginal, p=.07 | No, p=.80 | Yes, p=.01 |

Note: As in the original paper, legal cases are reported as medians, and public cases are reported as means, windsorized at 500. Legal comparisons are done with Wilcoxon rank sum tests, and public comparisons are done with t-tests.

### *B. Discussion: Judicious Replication*

In summary, our first set of replications can be best described as partially successful. In one case, the main effect replicated but the effect size decreased; in another, the main effect failed to replicate but the subsidiary temperature questions did; and the Sunstein results suggest that baselines have moved.

Appendix D, which examines these effects in granular detail with attention to individual differences, strongly suggests that these successes and failures have little to do with differences between platforms (since those differences were minor) or demographic weighting. Indeed, regression analysis suggests that differences between platforms are mostly driven by attentiveness: more attentive subjects (MTurk and the lab) produce results closer to the original, while less attentive subjects (SSI) create more noise and less similar results.

The failures to replicate are both puzzling but also perhaps predictable, particularly on a revisiting of the study materials. The materials that we showed subjects in all three studies evoke slightly off-kilter contexts for the contemporary reader.

- The apologies study describes a pedestrian/bike path through a local park. The study subjects in that case were almost all drawn from staff at a Midwestern University—presumably that scenario had resonance and

familiarity for them, but does not strike us as a universally-relatable context. The settlement questions themselves were also dissonant for slippery reasons. We do not have a sense of neighborly transactions in the original time and place of the study, but for our research team, the exchange of money between neighbors for personal injuries stemming from an accident seemed surprising or awkward.

- The materials in the Rachlinski study were similarly both straightforward enough and yet not what we would have chosen. Asking MTurk respondents to put themselves into the role of attorney, for example, is something we would normally (though not always) prefer to avoid, because it adds a layer of complexity for subjects who do not have experience with attorneys.
- The Sunstein study is most striking, of course. We do not talk about saving dolphins the way we used to, for example—it's just not the same salient exemplar of environmental advocacy. Similarly, given the Trump campaign's immigration rhetoric, the situation of migrant farm workers likely was a more salient example of injustice for those inclined toward that view.

One view of replication is that it mechanical; if your experiment is reproducible, you should be able to pick it up and put it down somewhere else, at a later time, with the same results. But psychology experiments are unlike the hard sciences in many ways, and one of these is the level of creative intuition brought to bear on the experimental operationalization. If we test a blood pressure medication, the hypothesis is that it reduces heart attacks, and the experiment is a direct test. If the hypothesis is that plaintiffs are more willing to settle than defendants, outside of a field experiment, the experiment requires a plausible narrative for which our subjects can engage, empathize, access the relevant intuitions, and report them. Identifying the relevant narrative—whether it is a vignette or a behavioral protocol in a lab setting—demands that the researcher mine her understanding of the world and make a set of judgments. Is this scenario plausible? Is it relatable for most people in my sample population? Is this amount of money a lot or a little? Is this the kind of thing that makes people angry?

We are not the first to understand this, but it bears repeating in this context specifically, because studies of legal behavior almost always draw on complex, contingent intuitions about the social and political moment. For example, early on in the project we had an idea that we ultimately discarded. We thought we'd like to try to mechanically replicate Janice Nadler's 2005 study on backlash to unfair laws, in which

she predicted that exposure to an unjust legal outcome would increase the likelihood of legal "flouting"—disregarding the legal rule when it seemed more fair to do so. Her stimuli included a story about the brutal murder of a young African-American girl by a privileged white adolescent, and another story involving a three-strikes law. The fifteen years since she first ran those studies have seen Barack Obama, Donald Trump, Black Lives Matter, and *Ewing vs. California*. Our cultural narratives around race, gender, policing, and sentencing have all seen significant shifts since then. A research team trying to test her hypothesis afresh in 2017 would never use those stimuli—they are dated and as such all the wrong details are salient. They press the wrong buttons. Our sense is that other vignettes may not be quite as jarringly dissonant, but that a replication team ought to take its own measure of the materials and think as systematically as possible about the costs and benefits of the original stimuli in light of the hypothesis being tested.

Our initial efforts into replication changed our understanding of what replication entails, and what it means. The replication efforts reported in the previous section were quite rote. We used the author's materials, made very few updates, and the results were consistent with the original effects, but clearly not perfect replications of the target phenomena. We decided to give ourselves one additional chance to replicate, taking into account the lessons from the first round.

We focused our attention on Rachlinski's 2002 experiment. In that study, he asked participants to imagine themselves as attorneys representing owners of leisure property (a bed and breakfast and a vacation home). We rewrote the scenario to put subjects in the position of the owners of small residential rental properties. We also used lower stakes--$25,000 rather than $200,000. Bed and breakfast inns and country vacation homes seemed associated with a particularly narrow slice of the American experience, one not especially accessible to most of our respondents. In an alternate scenario, we described a dispute between two large videogame firms—not the kind of thing our subjects would normally have participated in, but, given their youth and level of online activity, a narrative that would feel engaging and cognizable. [68]

In all cases, subjects were assigned to read that they were in the position of the plaintiff or of the defendant, and then given high or low odds of the plaintiff's success at trial. As in the original study, they answered the question: would you settle for an amount roughly equal to the expected value of the plaintiff's claim. We first ran both scenarios on MTurk. We then added, as a comparison, a recruitment from Prolific Academic, a highly-touted new platform which approximates MTurk's utility but was built to serve

[68] See Appendix A for details.

academic audiences and thus avoids some of its vices.[69] Our view was to essentially compare MTurk with its actual competitors, rather than with a "representative" paid sample.

Every subject on MTurk read both scenarios in a single condition—plaintiff win, plaintiff lose, defendant win, and defendant lose. We reproduced the property study only on MTurk. The details of each run are shown in Table 8.

**Table 8. Cost and Demographics of MTurk and Prolific studies**

| | MTurk | Prolific |
|---|---|---|
| Run Date | 6/15/2017 | 6/15/2017 |
| Cost | $734 ($1.50 per person plus fees) | $370 ($1.00/person plus fees) |
| Number | 402 | 303 |
| Percent Male | 54% | 60% |
| Median Age | 34 | 31 |
| Race | 85% White; 7% Black; 5% Asian | 82% White; 6% Black; 10% Asian |

Table 9 shows the results, which were markedly similar across platforms. Predicted effects were statistically significant, and overall effect size was more similar to the original studies—differences between plaintiffs and defendants were between 25 and 35 percentage points depending on the probability of winning, the scenario, and the subject pool.

**Table 9. Results of Judicious Replication**

| | **MTurk Results** | | | **Prolific Results** | | |
|---|---|---|---|---|---|---|
| | **% Ps Settling** | **%Ds Settling** | **Statistical testing** | **%Ps Settling** | **%Ds Settling** | **Sig.** |
| Losing: Property | 91.8 | 66.2 | t(117)=3.99, p=.0001 | 90.8% | 63.1% | t(117)=4.25, p=.0001 |

[69] *See* Peer e*t al.*, supra note 47, and accompanying text.

| | | | | | | |
|---|---|---|---|---|---|---|
| Winning: Property | 94.9% | 63.6% | t(106)=5.15, p<.0001 | 93.4% | 62.7% | t(110)=4.87, p<.0001 |
| Losing: Trademark | 79.6% | 44.6% | t(115)=4.25, p<.0001 | | | |
| Winning: Trademark | 92.3% | 58.8% | t(69)=4.41, p<.0001 | | | |

For some reasons that are concrete, and others that are more difficult to articulate, when we looked at the original study materials they felt dated. Rachlinski's hypothesis, however, was quite straightforward and within the realm of our joint research experience. We used the same basic method as the original, a scenario study, but tested the plaintiff/defendant settlement prediction using the specific content that we judged reasonably likely to elicit the relevant intuitions, should the prediction be right. There are clearly contexts and versions of this study that would not produce the same effect; we do not know the boundary conditions of this phenomenon. But now, like Rachlinski, we have offered evidence that in some contexts, defendants and plaintiffs may have starkly different assessments of settlement deals based solely on the gain/loss frame.

## IV. SYNTHESIS

This paper has reported three replication studies, each of which produced mixed results. Within those replication studies, we also evaluated the differential responses of participants recruited from three subject samples. We did not find dramatic differences in responses by sample, though we did find, consistent with the literature previously reviewed, that MTurk participants were more attentive to study materials.[70]

By contrast, we found that participants from SSI were markedly inattentive, and the results from that platform were the most divergent from both in person and other online samples. We note that it's understood that SSI has provided subjects for the in-house subject pool at Qualtrics, the popular survey platform which also offers researchers a source of subjects. This result suggests that researchers might take care in using these fee for service subject recruitment platforms when subject motivation is particularly obscure.

[70] *See* Appendix D for details.

Given the attentiveness and reliability of our MTurk respondents (along with the similar Prolific Academic sample), together with the now quite-large body of literature from cognate fields on the replicability of treatment effects using MTurk, we would urge peer review publications in our field to put aside some of the latent skepticism that they may have held for that platform, at the very least in comparison to paid survey firms.[71]

In this final section of the paper, we use both the results as well as our experience as researchers during this process, to draw some inferences and some prescriptions for the next stage of law and psychology research.

### *A. Replicating Across Time*

The three experiments we tried to replicate here were run between the mid 1990s and the early 2000s, meaning that the materials we used were originally drafted between 15 and 20 years ago. To speak very broadly: some legal facts are clearly specific to a particular time—the three-strikes rule, for example. Others are clearly less sensitive to temporal or cultural context, like a story about a dispute over a property line. And then there is a huge space in the middle, where it is less obvious how the temporal context might change the meaning or valence of a particular fact. This proposition—cultural events have differing meaning and salience over time—borders on a truism, but it is worth pointing out that it has been the subject of serious empirical inquiry.

A recent meta-study of replication by Jay van Bavel and colleagues, who asked coders to rate a list of studies for "contextual sensitivity."[72] Coders were blind to replication results, but their ratings were highly predictive of replication—the more contextually sensitive a given topic was rated, the less likely the replication attempt was successful. The specific point about temporal context is made vividly by Wolfgang Stroebe and Fritz Strack in a response to the replication debate in *Perspectives on Psychological Science*:

> "Let us illustrate this point with some classic social psychological experiments. In their study of the effect of the severity of initiation to a group on liking for that

[71] We make these claims about MTurk and attentiveness with appropriate caution, understanding that attentiveness is not a substitute for representativeness. As others have pointed out, attentiveness as measured by screener questions is itself associated with particular demographic and ideological traits. With MTurk, inattentive subjects have not selected into the platform. With SSI and similar platforms, the subjects agree to participate and then prove inattentive. Our view is that the former presents fewer interpretation challenges for researchers, but this is not an uncontroversial position. *See. e.g.,* Adam Berinsky et al., *Separating the Shirkers from the Workers? Making Sure Survey Respondents Pay Attention on Self-Administered Surveys*, 58 AM. J. POL. SCI. 739 (2014).

[72] Jay Van Bavel, Petere Mende-Siedlecki, William Brady & Diego Reinero, *Contextual Sensitivity in Scientific Reproducibility*, 113 PNAS 6454 (2016).

> group, Aronson and Mills (1959) operationalized the severe initiation by having female participants read aloud "12 obscene words, e.g., fuck, cock, and screw" as well as "two vivid descriptions of sexual activity from contemporary novels." If repeated with today's female students, this manipulation might trigger amusement rather than embarrassment. Similarly, it is likely that a researcher who tried to induce fear about toothbrushing in high school students by telling them that improper care of their teeth might result in "cancer, paralysis or secondary diseases" (Janis & Feshbach, 1953) might arouse disbelief rather than fear."[73]

Our contention at this point is that psychology of law is much more likely to fall into the latter category—heavily mediated by individual experiences with the particular culture and legal system in which they live—than the former. In fact, there is probably a deeper set of questions, about the nature of law and psychology as a field, that we would have to answer in order to do justice to this topic. Law and psychology as a field or sub-field tends to include two sets of inquiries. One is applications of basic findings—the endowment effect, for example—to legal contexts. The other is specifically about the psychology of legal relationships and choices, where researchers begin with a proposition that law has a specific and salient meaning along social, cultural, emotional, and political dimensions.

For studies in the latter category, particularly, it will be quite difficult to replicate any study with precision, or to even speak coherently about 'precision' and 'fidelity'. As Stroebe and Strack have warned, the peril of direct replication efforts is that the "result will be a reproducibility coefficient that will not be greatly informative, because of justified doubts about whether the 'exact' replications succeeded in replicating the theoretical conditions realized in the original research."[74]

*B. Research Norms in Law and Psychology*

Implicit in our account is that the "replication crisis" in psychology has been overhyped and undertheorized. We have also claimed (and we think shown) that MTurk is not an unreliable subject pool – or at least not unreliable in ways distinct from other available platforms. With that said, there is a concern about MTurk and other cheap online samples which is rarely articulated, but which is both deeply felt and legitimate. Online samples are extraordinarily easy to use, and available to anyone. Any given researcher can quickly and easily run multiple similar studies and choose to publicize only the one that produces significant results—the file drawer problem. But we think

---

[73] Wolfgang Stroebe & Fritz Strack, *The Alleged Crisis and the Illusion of Exact Replication*, 9 Persp. Psych. Sci. 59 (2014).
[74] *Id.*

more important is that open online platforms remove a traditional barrier to entry. The challenge of recruiting subjects served as a filter. That filter is now gone. In law, this is an especially important change, because the gatekeepers at the publication side are often, though of course not always, students. And the audience for experimental work in law and psychology is law faculties, not psychologists. This means that it is possible for work to get published and publicized without ever getting reviewed by an expert in the methodological field.

We cannot easily dismiss such concerns, nor should we. Researchers in law and psychology ought to be concerned about the overall legitimacy of their field, especially to the extent that they hope to have impact beyond the pages of law journals. For that reason, we advance the following steps–all borrowed from more mature, cognate, disciplines—which if adopted by our peers, would do much to assuage the growing concerns about experimental legal studies.

### 1. Internal Replication

*Use Multiple Scenarios to Elicit the Same Mechanism*: Single experiment papers ought generally to be viewed with some suspicion: researchers ought to show that the same results appear in response to different stimuli. Thus, we take it as good news that Rachlinski's results replicated in response to two very different set of legal prompts, one sounding in trademark and the other in real property.

*Multiple Studies:* Most experimental tests of hypotheses in psychology are limited by the constraints of any single design; conversely, most psychological phenomena are best described by a series of studies that build on one another. It is traditional in psychology for researchers to publish 3 to 6 studies per paper. This practice serves as a kind of miniature replication process; each study is different, but each is ultimately offering support for a broader hypothesis.

*Exact Replication on Multiple Platforms*: Readers would feel more confident about law and psychology findings if they were assured they weren't the product of the idiosyncrasies of a particular subject pool, however internally reliable. When feasible, exact replications with alternate subject pools are not only reassuring but often quite informative.[75] A study that showed that both business school students and MTurk participants made similar mistakes in financial decision-making, for example, is more

[75] *See, e.g.,* Joseph Simmons, Leif Nelson and Uri Simonsohn, *False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant*, 22 PSYCH. SCI. 1359 (2011) (arguing for exact, not just conceptual, replication of results as a means of combatting false-positive publications).

than just evidence that the mistake exists—it also says something interesting about shared cognitive limits. We note that it is now possible for outside researchers to purchase time on in person university samples, and that the cost of such samples is not prohibitive. Obviously, as sample sizes increase, alternatives to MTurk become more difficult to arrange, since that platforms' current major advantage is that it scales well.

### 2. Independent Reviewability and Reproducibility

*Making Study Materials Publicly Available:* The best way to evaluate a study is to understand its methods. This means understanding what subjects did and saw and how they recorded their reactions and what was measured. Authors and journals can make work more easily reviewable, not to mention comprehensible, by including all of their materials, in the text or in an appendix. This is crucial whether the study is a lab game or a scenario study—the reader needs to have a fine-grained sense of what the subjects experienced in order to evaluate what inferences can be reasonably drawn from the results. John Iaonnidis has advocated for allowing or encouraging authors to publish a "fulsome description of methods" to make replication possible and helps establish research standards.[76]

*Housing Data Online*: Current law review practices with respect to data and study materials are erratic at best. In this way, law is significantly behind disciplines like economics and psychology. Journals ought to require that authors deposit the full text of the experimental materials, as well as the resulting data and scripts enabling replication. This is a simple, pragmatic reform with a technological solution that surprisingly has not be widely adopted, even though there are "hundreds of data repositories available for data storage."[77]

*Piloting Records and Pre-registration*: In psychology, as in all empirical disciplines, norms of research and reporting have to balance flexibility and creativity in the exploration phase with rigor and integrity in the testing phase. The line between these two phases is often ignored or at least blurred. Some fields have called for pre-registration of all experimental tests, in order to bind researchers to the masts of their original data plans, and to make the file drawer open to the public.[78] However, pre-registration is arguably costly; many researchers worry that pre-registration and pre-analysis plans

---

[76] John Iaonnidis, Marcus Munafo, Paolo Fusar-Poli, Brian Nosek & Sean David, *Publication and Other Reporting Biases in Cognitive Sciences: Detection, Prevalence and Prevention*, 15 TRENDS IN COG. SCI. 235 (2014).

[77] *Id.* at 239.

[78] *See, e.g.,* Marcatan Humphreys, Raul Sanchez de la Sierra, and Peter van der Windt, *Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration*, 21 POLI. ANALYSIS 1 (2013).

"inhibit exploratory work."[79] Experimenting demands substantial exploration. We would suggest a practice of describing, even in broad strokes, the series of the pilots and pre-testing that the authors tried before they landed on the published studies.

* * *

We could go further, of course. More sophisticated ways to avoid bias—like perturbing data before analyzing it, as Robert MacCoun has suggested[80]—are excellent practices but might not be sufficiently simple to be widely adopted. Thus, we would start with these ground rules, which would tend, on net, to make it harder to do bad research and easier for readers to understand the limits of the findings that they've been presented.

## Appendix A: Survey Text

### 1. Robbennolt Study, MTurk version

Please imagine that on a recent evening, you were taking a walk in a local park on a walking/biking path. The weather was nice and you passed other neighbors doing the same. All of a sudden you looked up and saw a bicycle coming down the path, swerving to avoid a rock. Before you could jump out of the way, the bicycle rider collided with you. You fell, hit your head on the ground, and passed out. The next thing you knew, you woke up in the hospital. As it turned out, you were severely injured in the incident: you suffered a broken arm that required surgery so that doctors could insert pins to set the fracture, your ribs were bruised, and you received numerous severe bruises and scratches on your face and upper torso.

A number of your neighbors saw the accident. Several reported that the bicyclist was going too fast given the curves and the number of people out walking. One said the biker was reaching for water while riding very fast. Another said that the same bike almost hit another pedestrian further up on the path.

The bike rider turns out to be someone from the neighborhood named Will Nathanson. You have been acquainted with Will for a couple of years—friendly enough to exchange hellos, but not close friends. <apologyIntroText> <apologyText>

The surgery on your arm left you with medical bills that are not covered by your medical insurance. In addition, you have had quite a bit of pain in your ribs, your arm is

---

[79] Lucas Coffman & Muriel Niederle, *Pre-Analysis Plans Have Limited Upside, Especially Where Replications are Feasible*, 29 J. ECON. PERSP. 81, 88 (2015).

[80] Robert MacCoun & Saul Perlmutter, *Hide Results to Seek the Truth: More Fields Should, Like Particle Physics, Adopt Blind Analysis to Thwart Bias*, 526 NATURE 187 (2015).

in a cast and the doctors tell you that it is likely that you will not be able to fully extend your arm ever again, and you have had to take sick leave from work.

<noApologyText> Because you don't know how you should handle the situation, you talk to your friend who is a real estate lawyer. She tells you that one option would be to bring a lawsuit seeking compensation for the medical bills, lost time at work, and pain and suffering. She says that she can recommend a good lawyer, and says that given the likely testimony of the witnesses to the accident, you might have a chance at winning. She notes, though, that winning in court is never guaranteed and that any lawsuit will come with high legal fees.

| Key, Condition | Text Inserted |
|---|---|
| apologyIntroText, Partial, Full | Because you were unconscious immediately following the collision, you didn't have a chance to talk with Will at that time. However, while you were at the hospital, Will left a voicemail for you on your phone: "Hi. This is Will Nathanson. We collided on the bike path earlier. |
| apologyText, Partial | I am sorry if you were hurt. I really hope that you feel better soon. Bye." |
| apologyText, Full | I am so sorry that you were hurt. The accident was all my fault. I was going too fast and not watching where I was going until it was too late. I wanted to make sure that you had my phone number. It is 555-3405. Please call me if you need anything. Bye." |
| noApologyText, None | Will has not contacted you and has not apologized for the incident. |

**2. Rachlinski Study, MTurk version**

**Plaintiff scenario [winning | losing]:**

Please imagine that you are a lawyer with a client in a land dispute. Your client, Tom Smith, owns a large property in Oregon, where he has a vacation home. During his last visit he was surprised to discover that the bed-and-breakfast inn on the neighboring property had been expanded—including a new set of rooms that were built on a 30 x 10

foot area of Tom's property. Tom filed a law suit to order Real Resorts, the company owning the inn, to remove the new rooms. Real Resorts has stipulated (agreed) that they erroneously built the new rooms on your client's land. Your client has stipulated that his land has suffered no real reduction in value resulting from the loss of use of that piece of the land.

Under Oregon law, as in most states, Real Resorts has clearly trespassed on Tom Smith's land, and is continuing to do so. The judge assigned to the case will have her choice of two different remedies: (1) order Real Resorts to remove the structure from your client's property, or (2) order your client to sell the corner of his property to Real Resorts for its market value, which is negligible ($50).

In previous contacts with the defendant, you have learned that if the judge orders Real Resorts to remove the buildings, that rather than tear them down, Real Resorts will offer your Tom $300,000 to purchase the corner of property they built on. Tom would definitely accept $300,000 for the land. This means that if the case goes to trial, your client will win either $300,000 or $50 for the piece of property, depending on the judge's decision. (Of course he keeps the rest of his land either way.)

You have consulted a senior partner with your firm who knows this judge. He has stated that she [is an | is not an] adamant defender of property rights, and [hates | loves] to order forced sales of land. Your colleague believes that there is a good chance the judge will rule [in your favor | against you]. He estimates the chance of winning an order against Real Resorts at about [70|30]%--so your client has a [70|30]% chance of winning $300,000, and a [30|70]% chance of winning $50.

Tom says that if he loses at trial, he will not appeal. Now it is one day before trial. The lawyer for Real Resorts calls and makes a settlement offer of $[210|90]},000. Rather than go to trial, your client can accept $[210|90],000. It is a non-negotiable, final offer.

**Defendant Scenario [winning | losing]:**

Please imagine that you are a lawyer with a client in a land dispute. Your client, Real Resorts Inc., owned by Rick Mattson, recently expanded one of its inns in Oregon. Unfortunately, due to a surveying error, a small but costly part of the new complex was inadvertently built on a 30 x 10 foot piece of property owned by a neighbor. The survey firm has filed for bankruptcy and there is no chance of getting any money from them for their error. The owner of the land, Tom Smith, filed a law suit to order Rick to remove the new rooms. Your client has stipulated (agreed) that Real Resorts erroneously built the

new rooms on Mr. Smith's land. Mr. Smith has stipulated that his land has suffered no real reduction in value resulting from the loss of use of that piece of the land.

Under Oregon law, as in most states, Real Resorts has clearly trespassed on Tom Smith's land, and is continuing to do so. The judge assigned to the case will have her choice of two different remedies: (1) order your client to remove the structure from Mr. Smith's property, or (2) order Mr. Smith to sell the corner of his property to your client for its market value, which is negligible ($50).

Rick has decided that if the judge orders him to remove the buildings, that he will offer Mr. Smith $300,000 for the corner of property that the inn is on, and that is an amount you are sure Mr. Smith will accept. This means that if the case goes to trial, your client will lose either $300,000 or $50 for the piece of property, depending on the judge's decision.

You have consulted a senior partner with your firm who knows this judge. He has stated that she [is not an | is an] adamant defender of property rights, and [likes | hates] to order forced sales of land. Your colleague believes that there is a good chance the judge will rule [in your favor | against you]. He estimates the chance of losing and facing an order to tear down the new building to be about [30|70]%--so your client has a [30|70]% chance of losing $300,000, and a [70|30]% chance of losing $50.

Rick says that if he loses at trial, he will not appeal. It is now one day before trial. The lawyer for Mr. Smith calls and makes a settlement offer of $[90|210],000. Rather than go to trial, your client can pay $[90|210],000. It is a non-negotiable, final offer.

### 3. Judicious Replication

| **Property Dispute:** | **Trademark Violation:** |
|---|---|
| You are the owner of two residential properties, one where you live and another that you rent out. The property that you rent out is a two-bedroom house on a small lot of land. You renovated it 5 years ago and added a second bathroom and a garage.<br><br>The house next door is also a rental, and its owners recently decided to put it on the market. | Pillar Corporation introduced a gaming console called the Stunt to the market in 2006. The Stunt is designed for in-home use, typically hooking up to a television set. Players purchase the games separately, which they can download or install onto the console to play. Pillar has trademarked the "Stunt" name. |

| | |
|---|---|
| Soon after their For Sale sign goes up, you receive a letter by certified mail. It informs you that in your neighbor's recent appraisal and survey of their property in preparation for sale, it came to light that about 150 square feet (a 10 x 15 foot area) of your garage and bathroom addition falls on the back right parcel of your neighbor's property. The letter, drafted on the letterhead of a local attorney, asks that you take down the encroachment. | Novan Inc. is a start-up technology company that recently introduced a hand-held device called the Stunt Stick. The Stunt Stick is a small device with motion sensors, cameras, and wireless connectivity that players use to create, execute, and evaluate competitive physical challenges. It is an electronic device but it is meant to facilitate in-person competitions across a variety of athletic domains.<br><br>Along with their attorneys, the CEOs and Chief Marketing Officers of each company are now in negotiations over the naming of the Stunt Stick. All parties agree that a court is likely to find that the Stunt Stick violates the Pillar Stunt trademark and they agree that a court would order Novan to stop using the name. Novan has agreed to stop, and Pillar has no objection to Novan's rebranding of the Stunt Stick as the Crash Stick. |

## Appendix B: Process for Removing Potential Duplicate Survey Takers on MTurk

A potential concern with online surveys is that people may open up multiple copies of the survey and that this can influence their responses. This includes destroying random assignment and participants seeing only one experimental condition. It could also create a demand effect where participants observe what the differences between conditions are and anticipate what the experimenters are looking for and answer accordingly. It removes naïveté and allows participants to reread text that other participants did not have access to. It is crucial to attempt to remove any participants who may have seen more than one version of the survey. The following is our best practices procedure for cleaning data for this purpose.

First, we create and use the qualification feature of MTurk. A unique Qualification Type is created for each study series. During pilot testing every participant will be added to this qualification. In the MTurk HIT creation settings, workers must meet the criteria that the qualification "has not been granted." HIT Visibility is set to "Hidden - Only Workers that meet my HIT Qualification requirements can see and preview my HITs."

In the Qualtrics survey, IP address logging is turned on. Immediately after the introductory consent page, before viewing any study content, there is a page that asks for the participant's MTurk ID which requires a mandatory response.

When a HIT is completed, we stop data collection on the Qualtrics survey, including closing out any partially completed survey responses, in order to compare these with the completed results. The dataset we download includes the partially completed surveys, which are flagged as not finished.

We compare submitted HIT worker IDs with IDs entered on the second page of the survey. We also compare IP addresses and IDs amongst all survey entries, whether from fully completed or partially completed surveys. We flag all survey data that have an ID and/or an IP address that was entered more than once. We also flag any data that did not have a matching completed HIT with the provided ID.

Finally, we remove all flagged data rows, so that all partially completed surveys, surveys not matching an MTurk HIT, and surveys with duplicate IDs or IP addresses are not in the final data set.

All of the MTurk worker IDs recorded during are added to the qualification. This includes the ID from every completed HIT, as well as any extra IDs that were manually entered by users, whether or not the survey was actually completed.

## Appendix C: MTurk Differences by Day

Below are summary statistics for each of the demographic variables we collected across the three studies we ran.

**Age**

| | N | Mean | Std. Dev. | Median | Min | Max |
|---|---|---|---|---|---|---|
| Weekday | 527 | 34.9 | 11.5 | 31 | 18 | 73 |
| Weekday Eve: | 539 | 37.8 | 11.9 | 35 | 18 | 78 |
| Weekend: | 539 | 37.9 | 12.5 | 35 | 18 | 77 |
| All: | 1605 | 36.9 | 12.0 | 34 | 18 | 78 |

**% Male**

| | N | Mean |
|---|---|---|
| Weekday | 526 | 53% |
| Weekday Eve: | 539 | 48% |
| Weekend: | 538 | 43% |
| All: | 1603 | 48% |

**Education**

| | N | % some college | % 4 year degree | Professional / grad degree |
|---|---|---|---|---|
| Weekday | 527 | 88% | 47% | 28% |
| Weekday Eve: | 539 | 90% | 52% | 33% |
| Weekend: | 539 | 89% | 51% | 36% |
| All: | 1605 | 89% | 50% | 33% |

**Household Income**

| | N | <50K | 50-100K | 100-150K | >150K |
|---|---|---|---|---|---|
| Weekday | 528 | 53% | 35% | 8% | 3% |
| Weekday Eve: | 539 | 51% | 37% | 10% | 2% |
| Weekend: | 539 | 52% | 40% | 7% | 2% |
| All: | 1606 | 52% | 37% | 8% | 2% |

**Employment**

| | N | Full-time | Part-time | Retired | Unemployed | Student |
|---|---|---|---|---|---|---|
| Weekday | 528 | 60% | 16% | 17% | 36% | 5% |
| Weekday Eve: | 539 | 61% | 17% | 21% | 34% | 4% |
| Weekend: | 539 | 55% | 22% | 21% | 35% | 4% |
| All: | 1606 | 59% | 18% | 20% | 35% | 4% |

## Appendix D: On Individual Differences and Platform Assignment

Generally speaking, law and psychology papers focus on main effects, not individual differences.[81] However, especially with a large number of subjects, the effects of such investigations can be informative, though care must be taken in interpreting results causally. [82] In this Appendix, we report regression results for two of the studies we have run to control both for subject identity (demographic characteristics) and subject source (platform).

### Robbenolt Study

We first begin with the dichotomous willingness to accept settlement variable, which, as the text above demonstrated, was not powerfully influenced by the conditional assignment. That result hold following an OLS regression, as demonstrated in Column 1 of Table 10, as the apology conditional assignment is not significant. Table 10, Column 2, further demonstrates that there are some significant differences between platforms in the likelihood to accept. Column 3 shows that there is no significant interaction effect between conditional assignment and platform.

[81] *See* Jeffrey J. Rachlinski, *Cognitive Errors, Individual Differences, and Paternalism*, 73 U. Chi. L. Rev. 207 (2006).

[82] *See* Jacob M. Montgomery et al, *How conditioning on post-treatment variables can ruin your experiment and what to do about it*, Working Paper, available at http://www.dartmouth.edu/~nyhan/post-treatment-bias.pdf (June 30, 2017).

| | 1. Conditional Assignment and Demographic Dummies | | 2. Conditional Assignment and Platform Assignment | | 3. Platform, Condition, Interaction Terms | |
|---|---|---|---|---|---|---|
| Part Apology[83] | -0.02 | (0.03) | -0.03 | (0.03) | | |
| Full Apology | 0.04 | (0.03) | 0.02 | (0.03) | | |
| Age | -0.004*** | (0.00) | -0.01*** | (0.00) | -0.01*** | (0.00) |
| Woman | 0.07*** | (0.03) | 0.06*** | (0.03) | 0.07*** | (0.02) |
| <50,00 Income | 0.03 | (0.03) | 0.06** | (0.03) | 0.06** | (0.03) |
| Employed | -0.00 | (0.03) | -0.01 | (0.03) | -0.01 | (0.03) |
| SSI (v. MTurk) | | | 0.04 | (0.03) | | |
| Booth (v. MTurk) | | | -0.08*** | (0.04) | | |
| No Apology v. Any | | | | | -0.02 | (0.05) |
| No Apology * | | | | | | |
| SSI | | | | | 0.02 | (0.07) |
| Booth | | | | | 0.04 | (0.07) |
| Constant | 0.73*** | (0.06) | 0.77*** | (0.07) | 0.76*** | (0.07) |
| $R^2$ | 0.03 | | 0.04 | | 0.04 | |
| *Observations* | 1,533 | | 1,533 | | 1,533 | |

Table 10: OLS regression with robust standard errors on willingness to settle; * are significant at the 0.10 level; ** at the 0.05 level; *** at the 0.01 level.

In unreported regressions, we replicate these basic findings using as an outcome variable the subjects' perception of the events' influence on their likelihood to take care in the future.

**Rachlinski Study**

We now repeat this individualized analysis with the original replication of Rachlinski's results. Table 5 in the main text, *supra*, illustrates that while we found differences for winning parties in MTurk and the Lab, we did not do so for SSI, and we identified significant differences for losing parties only on MTurk. Tables 11 reports the results of OLS models (with robust standard errors) on the accept variable for winning parties. Similar results for losing parties are available in unreported regressions.

[83] Baseline is no apology.

| | MTurk | | SSI | | Booth | |
|---|---|---|---|---|---|---|
| Plaintiff v. Defendant | 0.15*** | (0.06) | 0.05 | (0.05) | 0.13** | (0.06) |
| Age | 0.01** | (0.00) | 0.003** | (0.00) | 0.00 | (0.02) |
| Woman | -0.02 | (0.06) | -0.02 | (0.05) | -0.05 | (0.07) |
| Employed | 0.00 | (0.07) | 0.03 | (0.06) | -0.04 | (0.06) |
| <50,00 Income | -0.12** | (0.05) | -0.11** | (0.05) | -0.01 | (0.08) |
| Constant | 0.55*** | (0.12) | 0.66*** | (0.11) | 0.66*** | (0.13) |
| R-2 | 0.09 | | 0.04 | | 0.03 | |
| Observations | 224 | | 324 | | 215 | |

Table 11: OLS regression with robust standard errors. Winning Parties; * are significant at the 0.10 level; ** at the 0.05 level; *** at the 0.01 level.

Table 12 combines data across platforms and illustrates the results of a logit regression that controls for subject source.

| | Subjects in Losing Conditions | | Subjects in Winning Conditions | |
|---|---|---|---|---|
| Age | 0.00 | (0.00) | 0.03*** | (0.00) |
| Woman | 0.08*** | (0.03) | -0.03 | (0.03) |
| Employed | -0.01 | (0.03) | -0.00 | (0.03) |
| <50,00 Income | 004 | (0.03) | -0.10*** | (0.03) |
| Plaintiff v. Defendant | 0.14*** | (0.05) | 0.16*** | (0.06) |
| SSI (v. MTurk) | -0.06 | (0.05) | 0.04 | (0.06) |
| Booth (v. MTurk) | 0.07 | (0.05) | 0.00 | (0.07) |
| Experimental Condition * SSI | -0.10 | (0.07) | -0.11 | (0.07) |
| Experimental Condition * Booth | -0.09 | (0.07) | -0.01 | (0.08) |
| Constant | 0.67*** | (0.06) | 0.63*** | (0.08) |
| R-2 | 0.03 | | 0.05 | |
| Observations | 777 | | 763 | |

Table 12: OLS regression with robust standard errors; * are significant at the 0.10 level; ** at the 0.05 level; *** at the 0.01 level.

Here, though the base rates between platforms are substantive different, there is no significant evidence of interaction between platforms and treatments. With a larger number of subjects, we confirm the effect of the experimental manipulation in *both* conditions in the directions anticipated. (The utility of using larger samples with smaller than previously described effect sizes is, of course, the precise point of the criticism of the replication project we described above.)