2-2023

Metaresearch, Psychology, and Law: A Case Study on Implicit Bias

Jason Chin

Alexander Holcombe
University of Sydney

Kathryn Zeiler
Boston University School of Law

Patrick Forscher

Ann Guo

METARESEARCH, PSYCHOLOGY, AND LAW: A CASE STUDY ON IMPLICIT BIAS

JASON M. CHIN,[*] ALEX O. HOLCOMBE,[**] KATHRYN ZEILER,[***] PATRICK S. FORSCHER,[****] & ANN GUO[*****]

ABSTRACT

When can scientific findings from experimental psychology be confidently applied to legal issues? And when applications have clear limits, do legal commentators readily acknowledge them? To address these questions, we survey recent findings from an emerging field of research on research (i.e., metaresearch). We find that many aspects of experimental psychology's research and reporting practices threaten the validity and generalizability of legally relevant research findings, including those relied on by courts and policy-setting bodies. As a case study, we appraise the empirical claims relied on by commentators claiming that implicit bias deeply affects legal proceedings and practices, and that training can be used to reduce that bias. We find that these claims carry many indicia of unreliability. Only limited evidence indicates that interventions designed to reduce prejudicial behavior through implicit bias training are effective, and the research area shows many signs of publication bias. To examine whether law journal articles are acknowledging these limits, we collected a sample of 100 law journal articles mentioning "implicit bias training" published from 2017-2021. Of those 100 articles, 58 recommend implicit bias training and only 8 of those 58 express any skepticism about its effectiveness. Overall, only 19 articles express skepticism about implicit bias training. We end with recommendations for law journal authors, researchers, and practitioners towards more credible application of psychology findings in law research and policy. Our focus is on how empirical research can be best used to solve our most important social issues including racism.

[*] Australian National University College of Law.
** University of Sydney School of Psychology.
*** Boston University School of Law
**** Busara Center for Behavioral Economics.
***** Faculty of Medicine Dentistry and Health Sciences, University of Melbourne (MD candidate). Correspondence should be addressed to Jason M. Chin; E-mail: jason.chin@anu.edu.au.
We thank participants of a Hebrew University faculty workshop for comments and suggestions. Alexis O'Hanlon (B.U. Law) provided excellent research assistance. Conflicts of interest statement: The authors declare no conflicts of interest. Funding statement: This article was partially funded through the Denison Scholar program at the University of Sydney. Boston University School of Law funded research assistance.

TABLE OF CONTENTS

INTRODUCTION

Law and psychology are natural partners.[1] Thinking, feeling, and behaving are at the heart of many disputes that legal systems seek to regulate. Accordingly, the promise of psychologically-informed law and policy often captures the attention of legal scholars,[2] policymakers,[3] and courts.[4] Although psychology research has much to offer law, the limits of its usefulness are equally important to understand and recognize. Yet, those limits appear to have received far less attention.

Disagreements about whether psychology findings are sufficiently well researched, tested, and agreed upon to inform the legal system date back to the beginnings of legal psychology.[5] A little over a century ago, John Henry Wigmore, a scholar and teacher of evidence, wrote a scathing critique of the work of Hugo Münsterberg and the fledgling field of legal psychology.[6] Münsterberg, who has been referred to as the "father of legal psychology,"[7] was a psychologist at Harvard and a highly public figure at the time for his involvement in several notorious legal cases.[8]

Wigmore's article was in large part inspired by Münsterberg's highly-publicized work in criminal cases across the United States.[9] For example, in a 1907 murder trial, Münsterberg administered a battery of psychological tests, such as timed word association tasks, on a key witness. He concluded that the witness was not intentionally lying. Münsterberg leveraged this fieldwork to advocate for a larger role for psychology in law. For instance, in a book of reflections on his role in various legal matters,

---

[1] For the sake of readability, this manuscript will collapse related and overlapping fields of law and psychology, forensic psychology, and correctional psychology, referring to them generally as "legal psychology" and "law and psychology." In using those terms, we refer to any application of psychology to law and policy. This admittedly elides many subtleties but does so for the sake of clarity. On the distinctions between some of those fields, *see generally* Tess M. S. Neal, *Forensic Psychology and Correctional Psychology: Distinct but Related Subfields of Psychological Science and Practice*, 73 AM. PSYCH. 651 (2018).

[2] *See generally, e.g.*, John H. Wigmore, *Professor Muensterberg and the Psychology of Testimony*: *Being A Report of the Case of Cokestone v. Muensterberg*, 3 ILL. L. REV. 399 (1909); Jerry Kang, Mark Bennett, Devon Carbado, Pam Casey & Justin Levinson, *Implicit Bias in the Courtroom*, 59 UCLA L. REV. 1124 (2012).

[3] *See generally, e.g.*, Jason M. Chin, Malgorzata Lagisz & Shinichi Nakagawa, *Where Is the Evidence in Evidence-Based Law Reform?*, 45 UNSW L. J. 1124 (2022).

[4] *See generally, e.g.*, State v. Henderson, 27 A.3d 872, 918–19 (N.J. 2011); Tess M. S. Neal, Christopher Slobogin, Michael J. Saks, David L. Faigman & Kurt F. Geisinger, *Psychological Assessments in Legal Contexts: Are Courts Keeping "Junk Science" Out of the Courtroom?*, 20 PSYCH. SCI. IN THE PUB. INT. 135 (2019).

[5] *See generally* Wigmore, *supra* note 2; Eilis S. Magner, *Wigmore Confronts Munsterberg: Present Relevance of a Classic Debate*, 13 SYDNEY L. REV. 121, 121 (1991); Brian H. Bornstein & Steven D. Penrod, *Hugo Who? G. F. Arnold's Alternative Early Approach to Psychology and Law*, 22 APPLIED COGNITIVE PSYCH. 759, 765 (2008).

[6] *See* Wigmore, *supra* note 2, at 407-434. The literature provides various spellings of "Münsterberg." We use the version with the German umlaut unless quoting other work.

[7] Magner, *supra* note 5, at 121.

[8] *Id*. at 122-26.

[9] Merle J. Moskowitz, *Hugo Munsterberg: A Study in the History of Applied Psychology*, 32 AM. PSYCH. 824, 831-833.

Münsterberg argued that there was little about a trial that could not be improved through the involvement of psychology research:

> **There is thus really no doubt that experimental psychology can furnish amply everything which the court demands**: it can register objectively the symptoms of the emotions and . . . it can trace emotions through involuntary movements, breathing, pulse, and so on, where ordinary observation fails entirely.[10]

Given this sort of claim, Münsterberg predictably attracted critics, prominent among them Wigmore in his 1908 article, "Professor Muensterberg and the Psychology of Testimony being a Report of the Case of Cokestone v Muensterberg."[11] That article, although satirical, raised questions that still endure.[12] For instance, Wigmore asked "Does this new method give a safe criterion for testing the individual witness?"[13] This challenge of applying group level research to individuals has been the focus of a great deal of recent legal-scientific research.[14] Wigmore also highlighted concerns with whether psychological testing was sufficiently "exact and precise"[15] and whether psychological claims had reached general acceptance in the field,[16] both also subjects of modern study.[17] And there also remains the problem demonstrated by Münsterberg's own testimony of psychology experts providing exaggerated and unsupportable claims to advance a party's interest.[18]

In retrospect, it is unsurprising that a great deal of psychology research would fall short at the time of Wigmore's missive. The practice of conducting systematic empirical studies to test research questions in psychology was only a few decades old.[19] The succeeding 110 years have seen a great deal of empirical research and theoretical development in legal psychology.[20] Alongside new findings, researchers have also developed new frameworks to understand when an empirical psychology claim has withstood critical scrutiny[21] and when it is more versus less likely to

---

[10] *Id*. at 832 [**emphasis added**].
[11] Wigmore, *supra* note 2.
[12] *See* Magner, *supra* note 5, at 121-22.
[13] Wigmore, *supra* note 2, at 421 [**emphasis added**].
[14] *See, e.g.*, David L. Faigman, John Monahan & Christopher Slobogin, *Group to Individual (G2i) Inference in Scientific Expert Testimony*, 81 UNIV. CHI. L. REV. 417, 417-418 (2014); Julia M. Haaf & Jeffrey N. Rouder, *Some Do and Some Don't? Accounting for Variability of Individual Difference Structures*, 26 PSYCHONOMIC BULL. & REV. 772, 772 (2019).
[15] Wigmore, *supra* note 2, at 401, 420.
[16] *Id*. at 425 ("Are its results yet even agreed upon?").
[17] On precision, *see infra* Part I.A. On general acceptance, *see infra* Part I.E.
[18] In R v. Pearce, 2014 MBCA 70, 14, the psychologist Jordan Peterson served as an expert witness and claimed he had invented an "unfakeable" personality test.
[19] *See* Magner, *supra* note 5, at 133.
[20] *See generally,* RUSSIL DURRANT, AN INTRODUCTION TO CRIMINAL PSYCHOLOGY (2d ed. 2017).
[21] *See e.g.,* Daniel Lakens, *The Value of Preregistration for Psychological Science: A Conceptual Analysis*, 62 JAPANESE PSYCH. REV. 221, 221 (2019) (noting that preregistration allows others to evaluate hypothesis tests' strength).

generalize to a population of interest.[22] Many of these insights were inspired by what has been called a "credibility revolution"[23] in psychology that responded to the failure of many prominent studies to replicate (sometimes referred to as a "crisis").[24] Large-scale replication studies, error detection, and general research on research practices are ongoing in a field referred to as "metaresearch"[25] or "metascience."[26] These developments–those focused on the study of research methods themselves–are the focus of the present article.

Yet, legal scholars do not regularly acknowledge that psychological evidence can be affected by problems that undermine its credibility, nor do they often reference the burgeoning metaresearch that studies and sometimes quantifies those problems.[27] This opens up the possibility that legal scholars, policymakers and judges are using psychology findings as rhetorical support for pre-existing policy preferences, without regard to the underlying strength of evidence.[28] We seek to address this problem by evaluating the limits of psychology research applied to law and policy. We go on to systematically examine whether law journal article authors have adequately acknowledged the limits of psychology research in one particular area–studies that explore the effectiveness of interventions aimed at reducing implicit bias. We find that acknowledgement is lacking and suggest several avenues towards developing a better approach to applying psychology research to law.

This article proceeds as follows. In Part I, we examine limits of the application of psychology research to legal issues. As noted, we focus on findings and analysis from the last decade of metaresearch in psychology and statistics. This includes research exploring the effects of using research designs with small sample sizes and undisclosed flexibility in analysis and

---

[22] Joseph Henrich, Steven J. Heine & Ara Norenzayan, *The Weirdest People in the World?*, 33 BEHAV. & BRAIN SCIS. 61, 61 (2010).

[23] Simine Vazire, *Implications of the Credibility Revolution for Productivity, Creativity, and Progress*, 13 PERSP. PSYCH. SCI. 411, 411 (2018). Note that economists have used "credibility revolution" to refer to improvements in the reliability of causal inference through better research design. Joshua D. Angrist and Jorn-Steffen Pischke, *The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics*, 24 J. ECON. PERSPEC. 3 (2010).

[24] *See* Leif D. Nelson, Joseph Simmons & Uri Simonsohn, *Psychology's Renaissance*, 69 ANN. REV. PSYCH. 511, 512 (2018) ("Many have been referring to this period as psychology's 'replication crisis.'").

[25] *See generally* Tom E. Hardwicke, Stylianos Serghiou, Perrine Janiaud, Valentin Danchev, Sophia Crüwell, Steven N. Goodman & John P. A. Ioannidis, *Calibrating the Scientific Ecosystem Through Meta-Research*, 7 ANNUAL REV. STATS. & ITS APPLICATION 11 (2020).

[26] Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware & John P. A. Ioannidis, *A Manifesto for Reproducible Science*, 1 NATURE HUM. BEHAV. 1, 1 (2017) ("The field of metascience — the scientific study of science itself — is flourishing…").

[27] *See infra* Part III.

[28] For earlier concerns about lack of warnings about potential limits of psychology research, *see Phoebe Ellsworth: Truth and Advocacy*, ASS'N PSYCH. SCI. (June 8, 2016), https://www.psychologicalscience.org/video/phoebe-ellsworth-truth-and-advocacy.html.

data collection.[29] We also consider research studying threats to the generalizability of psychology findings,[30] such as psychology's over-reliance on college student samples and other samples of convenience.[31]

Part II then seeks to apply that knowledge to a specific set of psychology findings related to implicit bias, which was initially presented as having vast consequences for the legal system.[32] The term "implicit bias" is meant to capture the possibility that people may act in certain ways due to "automatically activated associations about social groups"[33] that, according to some accounts, evade conscious awareness.[34] Law professors and other legal commentators have rung warning bells about implicit bias in the legal system for years, claiming that it affects the behaviors of lawyers,[35] police,[36] employers,[37] judges,[38] mediators,[39] and jurors.[40] In response, some jurisdictions mandate that judges engage in training to reduce their implicit bias.[41] And commentators sometimes characterize the research supporting such training as remarkably robust in a way reminiscent of Münsterberg's claims:

> While experts may disagree about the role such research should play in employment litigation, **the dispute is not over the validity of the research findings themselves**. Specifically, **there is agreement on the last of these four findings: implicit bias can be counteracted, interrupted, or corrected to prevent or reduce its impact on employment decisions**.[42]

Implicit bias is a useful case study because the view of many legal commentators diverges sharply from a view guided by the findings

---

[29] Joseph P. Simmons, Leif D. Nelson & Uri Simonsohn, *False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant*, 22 PSYCH. SCI. 1359, 1360 (2011) (describing how they used computer simulations to understand the effect of four commonly used questionable research practices).
[30] *See generally* Tal Yarkoni, *The Generalizability Crisis*, 45 BEHAV. & BRAIN SCIS. 1 (2022).
[31] *See* Henrich et al., *supra* note 22, at 82-83 (concluding that populations from which samples are commonly drawn are "distinct outlier[s] vis-à-vis other global samples").
[32] Kang et al., *supra* note 2, at 1135-68.
[33] Patrick S. Forscher & Patricia G. Devine, *Knowledge-Based Interventions Are More Likely to Reduce Legal Disparities Than Are Implicit Bias Interventions*, *in* ENHANCING JUSTICE: REDUCING BIAS 303-17 (Sarah E. Redfield & Am. Bar Ass'n eds., 2017).
[34] Kang et al., *supra* note 2, at 1129. For a review, *see* Bertram Gawronski, *Six Lessons for a Cogent Science of Implicit Bias and Its Criticism*, 14 PERSP. PSYCH. SCI. 574, 575-78 (2019).
[35] Kang et al., *supra* note 2, at 1167-68.
[36] *See generally* Megan Quattlebaum, *Let's Get Real: Behavioral Realism, Implicit Bias, and the Reasonable Police Officer*, 14 STAN. J. CIV. RTS. & CIV. LIBERTIES 1 (2018).
[37] Stephanie Bornstein, *Reckless Discrimination*, 105 CALIF. L. REV. 1055, 1055 (2017).
[38] Asha Amin, *Implicit Bias in the Courtroom and the Need for Reform*, 30 GEO. J. LEGAL ETHICS 575, 575 (2017).
[39] Carol Izumi, *Implicit Bias and Prejudice in Mediation*, 70 SMU L. REV. 681, 681 (2017).
[40] Jessica Blakemore, *Implicit Racial Bias and Public Defenders*, 29 GEO. J. LEGAL ETHICS 833, 833 (2016).
[41] State v. Plain, 898 N.W.2d 801, 841 (Iowa 2017); State v. Rashad, 484 S.W.3d 849, 860 (Mo. Ct. App. 2016) (Van Amburg, C.J., concurring).
[42] Bornstein, *supra* note 37, at 1096.

summarized in Part I. In other words, a great deal of implicit bias research carries hallmarks of untrustworthiness. In this way, it highlights the importance of caution among legal researchers and actors when evaluating that research and proposing reforms based on it. Given that such interventions are designed to reduce discriminatory behavior, it's vital that we understand the strength of the evidence base on which they are built. We demonstrate that the shaky evidence base requires consideration of alternative approaches if we hope to effectively address discriminatory behavior.[43]

Accordingly, Part III examines how scientific findings related to implicit bias are presented in law journals. Specifically, we conducted a systematic study of recent (2017-2022) mentions of "implicit bias training" in law journals indexed on HeinOnline, a widely used legal database. We chose mentions of training for two reasons. First, scant research finds that interventions aimed at reducing implicit bias are effective. Therefore, we would expect that if legal researchers accurately describe the research, they will be cautious when discussing implicit bias training. And second, whereas implicit bias itself is a relatively complicated construct that builds on a great deal of past research with varying degrees of research support over the past decade, studies on training are more straightforward. Specifically, there has never been strong support that interventions aimed at reducing implicit bias are effective, and so it would be especially problematic if legal scholars and practitioners do not acknowledge this limit.

In Part IV, we chart a path forward. What can psychologists, especially legal psychologists, do to make clearer the limits of their work? Here, we suggest tangible reforms, such as including "constraints on generality" statements in empirical work.[44] These statements express the researcher's view of the scope of their finding, such as the populations and contexts to which the finding can be applied. Such statements may assist legal researchers who do not have the expertise to know which results are likely to generalize, and which are not.[45] We also provide recommendations to legal scholars, lawyers, and others whose primary field is law.

Part V concludes on an optimistic note. Although the experience with implicit bias and many of the findings we discuss throughout this article are so tenuous that they are not yet ready to inform law and policy, psychology and law can be a fruitful partnership. Psychology experiments have, for example, provided useful demonstrations that human memory is

---

[43] These alternative approaches might include structural changes in the legal system and focusing our attention on alternatives to implicit bias training.

[44] Daniel J. Simons, Yuichi Shoda & D. Stephen Lindsay, *Constraints on Generality (COG): A Proposed Addition to All Empirical Papers*, 12 PERSPS. PSYCH. SCI. 1123, 1123 (2017).

[45] Ellsworth, *supra* note 28 ("We have to recognize that research on social issues gets communicated to a much broader audience than basic research. We can assume that our colleagues will be skeptical about our claims in the discussion section and they'll evaluate them in terms of what they read in the methods sections and the results that we actually got. People that are not scientists - legislators, judges, reporters, the public - are more likely to read only the introduction and discussion.").

reconstructive and that people, in some conditions, report rich false memories.[46] They have also provided evidence showing that traditional methods of taking eyewitness identifications[47] can increase misidentifications, likely contributing to wrongful convictions.[48] To leverage this capacity of psychology to inform crucial questions of law and policy, Part V lays out a metaresearch agenda for law and psychology. That is, we propose a research plan for how the field may begin to study its own methods more carefully so that it can improve and fulfil its potential.

I. The Discovery of limits on applying experimental psychology research to law and subsequent reforms to address them

Wigmore highlighted several limits in experimental psychology's ability to inform legal issues based on his subjective impression of the state of the field at the time.[49] In this Part, we draw on recent metaresearch-informed critiques of experimental psychology levied since Wigmore's time to develop a more sophisticated and modern understanding of why and in what contexts psychology is more or less fit for use. These critiques highlight not only fundamental methodological concerns over the credibility of experimental psychology findings but also the fact that the studies' authors fail to explicitly acknowledge such concerns. In fields that produce findings likely to be applied in policy reform, making readers aware of a study's limitations is especially important because those who wish to apply the findings may not have the expertise to critically appraise empirical research.

In an effort to illuminate modern credibility problems, we also review ongoing reforms that seek to make research more credible, such as data and analysis script disclosure requirements increasingly imposed by academic journals that publish empirical studies.[50] Disclosure of data and other research materials allows for others to audit and replicate the results, making such disclosure a potentially useful *indicium* of a study's reliability that users can use to gauge the general credibility of a finding.

---

[46] *See* Henry Otgaar, Mark L. Howe & Olivier Dodier, *What Can Expert Witnesses Reliably Say About Memory in the Courtroom?*, 3 FORENSIC SCI. INT. MIND & L. 100106, 2-3 (2022). However, the boundaries of these effects are often unknown. *See* Jason M. Chin & Tess M. S. Neal, *Further Caution is Required on What Memory Experts Can Reliably Say*, 3 FORENSIC SCI. INT. MIND & L. 1, 1 (2022).

[47] *See generally* Gary L. Wells, Margaret Bull Kovera, Amy Bradfield Douglass, Neil Brewer, Christian A. Meissner & John T. Wixted, *Policy and Procedure Recommendations for the Collection and Preservation of Eyewitness Identification Evidence*, 44(1) LAW & HUM. BEHAV. 3 (2020).

[48] Jeffrey Bellin, *The Evidence Rules that Convict the Innocent*, 106 CORNELL L. REV. 305, 325-30 (reviewing the role of eyewitness identifications in wrongful convictions).

[49] *See* Wigmore, *supra* note 2, at 404-34 (criticizing cases in which Münsterberg relied on research that had not yet reached scientific consensus).

[50] Edith Beerdsen, *Litigation Science After the Knowledge Crisis*, 106 CORNELL L. REV. 529, 529 (2021). Analysis scripts are lists of commands empirical study authors use to get statistical software packages to produce results that are reported in studies.

A. Uncertainty about the existence and magnitude of experimental psychology research findings

During the past decade, metaresearchers have cast doubt on whether published and seemingly well-supported findings in experimental psychology represent false positives and, to the extent reported effects exist, whether they are as large as previously believed.[51] In 2010, psychologists started noticing several warning signs about the state of their field.[52] The red flags included some cases of fraud but also several reports of failed replication attempts.[53] In other words, researchers attempted to follow the methods of published studies by collecting data using new subjects drawn from the same or similar populations, but they failed to find evidence similar to the original, generally accepted findings.[54] Some termed this a "replicability crisis."[55]

Two notable examples stand out. The first, what is known as "behavioral priming," is noteworthy because of its prominence in the legal-psychology literature. It once supported the notion that subtle exposure to cues in the environment could substantially affect behavior. As we discuss in more detail in Parts II and III, this effect underlies much of the theory for implicit bias in that it suggests behavior is influenced by processes that we are not aware of. In 2012, a series of studies failed to replicate the behavioral priming effect that exposing participants to old-age related words caused them to walk more slowly.[56] These failures eventually led one of behavioral priming's most well-known proponents, Daniel Kahneman, to declare that "behavioral priming research is effectively dead."[57]

---

[51] Open Science Collaboration (OSC), *Estimating the Reproducibility of Psychological Science*, 349 SCI. 943, 944 (2015); Richard A. Klein et al., *Investigating Variation in Replicability: A "Many Labs" Replication Project*, 45(3) SOC. PSYCH. 142, 150 (2014) ("The original studies produced underestimates of some effects…, and overestimates of other effects…. Two effects … did not replicate."); Richard A. Klein et al., *Many Labs 2: Investigating Variation in Replicability Across Samples and Settings*, 1 ADVANCES METHODS & PRAC. PSYCH. SCI. 443, 446 (2018) ("[W]e found that [of 28 replicated studies] 15 (54%) of the replications provided evidence of a statistically significant effect in the same direction as the original finding…. Seven (25%) of the replications yielded effect sizes larger than the original ones, and 21 (75%) yielded effect sizes smaller than the original ones."); Colin F. Camerer et al., *Evaluating Replicability of Laboratory Experiments in Economics*, 351 SCI. 1433, 1433 (2016) ("We found a significant effect in the same direction as in the original study for 11 [of 18] replications (61%); on average, the replicated effect size is 66% of the original.); Colin F. Camerer et al., *Evaluating the Replicability of Social Science Experiments in Nature and Science Between 2010 and 2015*, 2 NATURE HUM. BEHAV. 637, 637 (2018) (We find a significant effect in the same direction as the original study for 13 [of 21] (62%) studies, and the effect size of the replications is on average about 50% of the original effect size."). For definitions and information about common measures of effect sizes in psychology, see https://osf.io/3zgxe.

[52] Nelson, Simmons & Simonsohn, *supra* note 29 at 512-14.

[53] *Id*.

[54] *See* notes at *supra* note 51.

[55] Nelson et al., *supra* note 29 at 512.

[56] Stéphane Doyen, Olivier Klein, Cora-Lise Pichon & Axel Cleeremans, *Behavioral Priming: It's All in the Mind, but Whose Mind?*, 7 PLOS ONE e29081 (2012).

[57] EdgeCast, *Daniel Kahneman - Adversarial Collaboration*, APPLE PODCASTS, https://podcasts.apple.com/no/podcast/daniel-kahneman-adversarial-collaboration/id1451643895?i=1000552318928 (last accessed Feb. 6, 2023).

Alarms raised by these replication failures were heightened by the results of more systematic replication studies. Across several social scientific fields (and eventually bioscientific fields),[58] large multi-laboratory collaborations attempted to replicate studies published in widely-cited journals (e.g., 100 from psychology, 18 from experimental economics, 21 published in *Nature* and *Science*).[59] Troublingly, they found statistically significant results[60] consistent with the original findings only about 50% of the time, and the magnitude of the effects they found was about 60% smaller than effect sizes reported in the originals.[61] Such metaresearch findings demonstrate that many published experimental psychology studies overstate both the statistical significance and magnitude of reported effects.

Two types of fallout from these failed replication attempts are germane to understanding how best to use experimental psychology research to inform law.[62] First, they have inspired metaresearchers to explore why so many findings are not as credible as they may seem. Uncovered reasons include publication bias, researcher use of questionable research practices, and inadequate sample sizes, all of which we discuss below. Second, they have spurred researchers to reform their research and reporting processes. After discussing these newly discovered frailties in psychology research methods, we describe how the inadequate traditional safeguards of credibility are being replaced by promising new ones.

1. Publication bias

Ongoing metaresearch is shedding light on the degree to which psychology (among other fields) suffers from publication bias, which is the tendency of the published literature to contain predominantly studies supporting the stated hypothesis.[63] This tendency is likely driven by academic journals preferring flashy positive findings (at the $p = 0.05$ threshold) and researchers responding to that incentive structure.[64] In psychology specifically, publication bias is rampant. A recently estimated

---

[58] *See generally* Timothy M. Errington, Maya Mathur, Courtney K. Soderberg, Alexandria Denis, Nicole Perfito, Elizabeth Iorns & Brian A. Nosek, *Investigating the Replicability of Preclinical Cancer Biology*, 10 eLIFE 1 (2021).

[59] Open Science Collaboration, *supra* note 51, at 944; Camerer et al., (2016) *supra* note 51, at 1433; Camerer et al., (2018) *supra* note 51, at 637.

[60] Statistical significance is a way of quantifying the likelihood that what we observe is due to mere chance as opposed to some studied factor.

[61] Amanda Kvarven, Eirik Strømland & Magnus Johannesson, *Comparing Meta-Analyses and Preregistered Multiple-Laboratory Replication Projects*, 4 NATURE HUM. BEHAV. 423, 425 ("The average unweighted effect size is 0.163 for the 15 replication studies and 0.423 for the 15 meta-analysis studies, implying that the mean meta-analytic effect size is almost three times as large as the mean replication effect size."; (0.163-0.423) / 0.163=-0.62 or roughly 60% smaller) (2020); For a definition of effect sizes, *see* Daniël Lakens, *Effect Sizes*, IMPROVING YOUR STATISTICAL INFERENCES (2022), https://lakens.github.io/statistical_inferences/effectsize.html. We further discuss effect sizes *infra* Part I.A.4.

[62] For a fuller review *see* Hardwicke et al., *supra* note 25.

[63] Gerald J. Haeffel, *Psychology Needs to Get Tired of Winning*, 9 ROYAL SOC'Y OPEN SCI. 1, 1 (2022).

[64] *See* Jason M. Chin, *Psychological Science's Replicability Crisis and What It Means for Science in the Courtroom*, 20 PSYCH., PUB. POL'Y, & L. 225, 229 (2014).

95% of published studies report statistically significant findings[65] for the first-listed hypothesis in the paper.[66] Psychology studies with law and policy consequences also suffer from publication bias.[67] Bodies of research that suffer from publication bias are misleading. This is because the publication of only positive results, and not studies that failed to find a statistically significant effect, makes it seem as if a finding is robust and easily demonstrable across multiple contexts when that might not be the case.

Take, for example, research on "nudges," interventions that change the context or framing of a decision in order to affect the decision maker's choice. Nudges have fascinated legal commentators[68] for years and some governments have even implemented "nudge units" in response to this research.[69] Proponents of using nudges in public policy describe them as powerful[70] informal reviews of nudges in law journals often do not hint at the possibility of publication bias.[71] A 2022 meta-analysis of nudge studies,[72] however, found considerable evidence for publication bias.[73] The bias was so extreme that one metaresearch analysis concluded that "after correcting for this bias, no evidence remains that nudges are effective as tools for behaviour change."[74] Although this conclusion bucks the optimistic tone struck by nudge boosters, the study provides convincing

---

[65] *See* Hillel J. Bavli, *Credibility in Empirical Legal Analysis*, 87 BROOK. L. REV. 501, 507-09 (2022).

[66] Anne M. Scheel, Mitchell R. M. J. Schijen & Daniël Lakens, *An Excess of Positive Results: Comparing the Standard Psychology Literature with Registered Reports*, 4 ADVANCES METHODS & PRACS. PSYCH. SCI. 1, 6 (2021).

[67] Elizabeth Levy Paluck, Roni Porat, Chelsey S. Clark & Donald P. Green, *Prejudice Reduction: Progress and Challenges*, 72 ANN. REV. PSYCH. 533, 538 (2021); Patrick S. Forscher, Calvin K. Lai, Jordan R. Axt, Charles R. Ebersole, Michelle Herman, Patricia G. Devine & Brian A. Nosek, *A Meta-Analysis of Procedures to Change Implicit Measures*, 117 J. PERSONALITY & SOC. PSYCH. 522, 541 (2019).

[68] *See generally* RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS (2009); CASS R. SUNSTEIN & LUCIA A. REISCH, TRUSTING NUDGES: TOWARD A BILL OF RIGHTS FOR NUDGING (2019).

[69] Ben Quinn, *The 'nudge unit': the experts that became a prime UK export*, THE GUARDIAN (Nov. 11, 2018), https://www.theguardian.com/politics/2018/nov/10/nudge-unit-pushed-way-private-sector-behavioural-insights-team.

[70] THALER & SUNSTEIN, *supra* note 68, at 253 ("One of our main hopes is that an understanding […] the power of nudges, will lead others to think of creative ways to improve human lives in other domains.").

[71] *See e.g.,* Jacob Goldin, *Which Way to Nudge: Uncovering Preferences in the Behavioral Age*, 125 YALE L. J. 226, 240-242 (2015); Wendy Netter Epstein, *The Health Insurer Nudge*, 91 S. CAL. L. REV. 593, 633-636 (2018); Robert J. Landry II., *Credit Card Debt and Consumer Bankruptcy: Can We Nudge Our Way out*, 27 AM. BANKR. INST. L. REV. 139, 146-147 (2019).

[72] Stephanie Mertens, Mario Herberz, Ulf J. J. Hahnel & Tobias Brosch, *The Effectiveness of Nudging: A Meta-Analysis of Choice Architecture Interventions Across Behavioral Domains*, 119 PROC. NAT'L ACAD. SCIS. 1, 4 (2022).

[73] Specifically, they found that studies with larger sample sizes tended to find null or small effects, whereas the studies using smaller samples found larger effects. This pattern indicates publication bias because, among other reasons, one strategy for publishing positive results is to run several small studies and only publish the ones that find evidence for the effect. However, larger studies are more difficult to suppress because they are expensive to conduct.

[74] Maximilian Maier, František Bartoš, T. D. Stanley, David R. Shanks, Adam J. L. Harris & Eric-Jan Wagenmakers, *No Evidence for Nudging After Adjusting for Publication Bias*, 119(31) PROC. NAT'L ACAD. SCIS. 1, 2 (2022).

evidence that publication bias makes the findings supporting the efficacy of nudges appear much stronger than they may actually be.

Metaresearch on nudges, besides calling into question the empirical foundations of the tool's efficacy, highlights several problems for the applied field of legal psychology. For instance, it casts doubt on the usefulness of informal reviews of published experimental psychology research. This includes those in law journals[75] as well as those found in popular psychology books.[76] These reviews do not account for publication bias, and authors of the reviews do not appear to be proactively warning readers about those possibilities.[77]

2. Questionable research practices

A second reason to question the credibility of experimental psychology studies is the accumulating evidence of researchers' use of what have been termed "questionable research practices."[78] These practices exploit undisclosed flexibility in research methods and reporting to make findings seem superficially persuasive. Metaresearch has uncovered evidence that researchers, including experimental psychologists, regularly engage in a number of such practices.[79]

One questionable research practice is to measure a psychological process in multiple ways, but report only those measures that support the research hypothesis.[80] For instance, a marketing researcher who wants to know whether poor customer service makes customers angry might measure outward manifestations of anger in multiple ways, such as verbal aggression, scowls or other facial expressions, elevated voice, and so on. If all measures except one suggest that poor customer service does not make customers angry, the researcher may opt to report only the single measure that "worked." Selective reporting, of course, potentially misleads readers about whether people actually exhibit a particular response and how robust the phenomenon is.[81] Under typical statistical practices in psychology, if a

---

[75] *See e.g.,* Goldin, *supra* note 71, at 240-242; Epstein, *supra* note 71 at 633-636; Landry III, *supra* note 71 at 146-147.

[76] *See* THALER & SUNSTEIN, *supra* note 68 (containing an informal review of the psychological evidence for several nudges).

[77] We acknowledge the possibility that our informal search and review of law journal articles about nudges missed some that noted publication bias was a possibility.

[78] *See generally* Leslie K. John, George Loewenstein & Drazen Prelec, *Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling*, 23 PSYCH. SCI. 524 (2012). Questionable research practices are also known as "researcher degrees of freedom," *see* Simmons et al., *supra* note 29, at 1359, and exploitation of "analytic flexibility," *see* Beerdsen, *supra* note 49, at 533.

[79] Franca Agnoli, Jelte M. Wicherts, Coosje L. S. Veldkamp, Paolo Albiero & Roberto Cubelli, *Questionable Research Practices Among Italian Research Psychologists*, 12 PLOS ONE 1, 1 (2017); André L. A. Rabelo, Jéssica E. M. Farias, Maurício M. Sarmet, Teresa C. R. Joaquim, Raquel C. Hoersting, Luiz Victorino, João G. N. Modesto & Ronaldo Pilati, *Questionable Research Practices Among Brazilian Psychological Researchers: Results from a Replication Study and an International Comparison*, 55 INT'L J. PSYCH. 674, Table 1 (2020).

[80] A related questionable research practice is selective inclusion of data used to produce measures.

[81] Empirical legal researchers, themselves, are thought to engage in this questionable research practice. In fact, Bavli, *supra* note 65, at 507, refers to selective reporting as "the central problem in empirical legal research."

researcher measures 20 outcomes, the likelihood of obtaining at least one positive result that is, in reality, negative is approximately 64%.[82]

The consequences and prevalence of questionable research practices are troubling. Using just four of these practices can increase a study's false positive rate–the rate at which a significant difference is claimed when one does not exist in reality–from 5% all the way to 60%.[83] And, regarding prevalence, questionable research practices are widespread in psychology. Large surveys find that psychologists in the U.S.,[84] Italy,[85] and Brazil[86] report high levels of questionable research practice use. For example, 56% of psychologists in a U.S. survey said that they decided to collect more data after observing whether their current result was statistically significant, and 38% said they decided whether to exclude evidence after looking at its impact on the result.[87]

In addition to conducting surveys, metaresearchers also examine published studies themselves for signs of questionable research practices. One particularly easy-to-spot sign is a series of studies published in a single article all with *p*-values hovering just below or at 0.05.[88] Unfortunately, the *p*-value associated with a result, and specifically if it falls below 0.05, is a *de facto* prerequisite for publishing a finding in much of psychology.[89] As a result, when several *p*-values within a paper fall just below 0.05, it is a sign that the results may not be credible.[90] To better understand these patterns, metaresearchers have developed a technique to compare reported p-values (a "*p*-curve") to a distribution of *p*-values that would be expected in the absence of questionable research practices and publication bias.[91] The

---

[82] STEPHEN B. HULLEY, STEVEN R. CUMMINGS, WARREN S. BROWNER, DEBORAH G. GRADY & THOMAS B. NEWMAN, DESIGNING CLINICAL RESEARCH 59 (3d ed. 2007).

[83] *See, e.g.*, Simmons et al., *supra* note 29, at 1359-60 (discussing reasons behind non-malicious but self-serving researcher behavior).

[84] John et al., *supra* note 78, at 525 (Table 1) (surveying U.S. psychologists).

[85] Agnoli et al., *supra* note 79, at 1 (surveying Italian psychologists).

[86] Rabelo et al., *supra* note 79, at 675 (Table 1) (surveying Brazilian psychologists).

[87] John et al., *supra* note 78, at 525.

[88] Researchers use observed data to compute p-values, which are measures of the likelihood of observing a result (e.g., a difference in a behavioral measure between the treatment group and the control group) at least as extreme as the result actually observed assuming the null hypothesis—in psychology experiments, usually that the treatment has no effect—is true. See https://osf.io/s72de.

[89] *See* Marjan Bakker, Annette van Dijk & Jelte M. Wicherts, *The Rules of the Game Called Psychological Science*, 7 ASS'N PERSP. PSCY. SCI. 543, 543 (2012).

[90] Adam Altmejd, Anna Dreber, Eskil Forsell, Juergen Huber, Taisuke Imai, Magnus Johannesson, Michael Kirchler, Gideon Nave & Colin Camere, *Predicting the Replicability of Social Science Lab Experiments*, 14 PLOS ONE 1, 11 (2019) ("The statistical properties (p-value and effect size) of the original experiment are the most predictive.").

[91] Uri Simonsohn, Leif D. Nelson & Joseph P. Simmons, *P-Curve: A Key to the File-Drawer*, 143 J. EXPERIMENTAL PSYCH. GEN. 534, 535 (2014) ("As detailed below, only right-skewed p-curves, those with more low (e.g., .01s) than high (e.g., .04s) significant p values, are diagnostic of evidential value. P-curves that are not right-skewed suggest that the set of findings lacks evidential value, and p-curves that are left-skewed suggest the presence of intense [questionable research practices]."). Some have developed similar methods using z-curves, which improve comparisons under certain conditions. *See generally* Frantisek Bartos & Ulrich Schimmack, *Z-curve 2.0: Estimating Replication Rates and Discovery Rates*, 6 META-PSYCHOLOGY 2720 (2022).

authors of this technique provide an app that can be used to expose reported findings to their *p*-curve analysis.[92]

3. Small sample sizes and underpowered studies

A third contributor to uncertainty about the credibility of experimental psychology studies is the use of insufficient sample sizes to estimate effects.[93] Small samples increase the probability of reporting a false negative; that is, concluding that no effect exists when the effect, in fact, does exist. Here, an analogy to a telescope may help.[94] If a weak telescope is pointed at an area of the sky and does not make a planet visible to the user, a planet might be at those coordinates, but the telescope is not powerful enough to observe it. Similarly, studies that fail to detect something may fail not because there isn't a practically meaningful effect to be found. Rather, they may lack sufficiently powerful designs (e.g., a sufficient number of participants) to detect the effect. This is especially problematic in psychology, where effects sizes are often small[95] and thus require highly powered studies (e.g., many participants, precise measurements) to detect them.[96] The converse is also true. When studies purport to identify an effect, it is less likely to be a true one when the sample size is small than when it is large.[97]

Inadequate power has particularly destructive consequences when combined with publication bias because these two qualities together result in published studies *overestimating* the size of any effect they find due to a phenomenon known as the "winner's curse."[98] In short, the winner's curse refers to research on common value auctions, finding that the winner of an auction is also the party most likely to have overestimated the value of the good. Similarly, scientists who read and interpret only studies with significant results ($p < 0.05$) will develop an inflated sense of the effects they observe because they do not see the studies that find small effects. This can occur in fields plagued by publication bias. Exaggeration of effects will be especially drastic for small studies because, due to their "small telescopes" (to use our analogy from earlier), small studies will find a statistically significant difference only when they observe an effect that is

---

[92] *P*-CURVE.COM, http://p-curve.com/ (last accessed Feb. 7, 2023).
[93] *See generally* Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson & Marcus R. Munafò, *Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience*, 14 NATURE REVS. NEUROSCIENCE 365 (2013).
[94] Uri Simonsohn, *Small Telescopes: Detectability and the Evaluation of Replication Results*, 26 PSYCH. SCI. 559, 560 (2015) ("[For a test to locate a small effect] to be properly powered, it requires a sample size typically not feasible for psychology experiments.").
[95] As we discuss below, psychologists traditionally classify effects as "small," "medium," and "large," labels with little practical meaning. Lakens, *supra* note 61 ("A commonly used interpretation of Cohen's *d* is to refer to effect sizes as small (d = 0.2), medium (d = 0.5), and large (d = 0.8) based on benchmarks suggested by Cohen (1988). However, these values are arbitrary and should not be used.")
[96] *See* Simonsohn, *supra* note 94.
[97] Specifically, small samples reduce the positive predictive value or "PPV" of a study. That is, they reduce the probability that a positive finding reflects a true finding. *See* Button et al., *supra* note 93, at 366.
[98] *Id*. at 367.

extreme.[99] Because, in the presence of publication bias, small studies are especially prone to exaggerating the effects they find, whereas large studies are less prone to exaggeration, metaresearchers interpret across-study negative correlations between sample size and effect size as a publication bias indicator.

4. Inadequate traditional safeguards contrasted against promising new ones

Publication bias, questionable research practices and small samples contribute to uncertainty about the existence and size of effects reported in published experimental psychology studies. Recent metaresearch findings demonstrate that traditional safeguards designed to prevent unreliable research from becoming generally accepted clearly are not up to the task. The traditional peer review process, for example, is inadequate as a means of detecting questionable research practices because reviewers are unable to detect most of them in the reports that they evaluate (e.g., selectively unreported measures and conditions are, by definition, not reported). Such traditional safeguard failings have generated concern that published journal articles serve more as "advertising" for a study rather than veridical reports of what, in total, was actually planned, hypothesized, done and observed.[100]

In addition, peer reviewers often do not catch basic statistical errors that, if corrected, would change the interpretation of reported findings. Metaresearchers have identified a large number of statistical errors in published psychology studies.[101] Similarly, in medicine, researchers studying the effects of peer review training found that 162 participants who were not exposed to any training identified an average of 2.4 out of nine deliberately inserted major errors while 120 participants who took a self-taught peer review course identified 2.9 major errors on average.[102]

In response to concerns about questionable research practices and undetected errors, some journals and researchers have adopted more transparent and open practices. These practices, among other things, reduce the ability of researchers to leverage undisclosed flexibility in the research process. Eight such practices have been formalized in the Transparency and Openness (TOP) guidelines, which have been adopted, to one degree or

---

[99] *Id*. at 366-367

[100] Jonathan B. Buckheit & David L. Donoho, *Wavelab and Reproducible Research*, *in* WAVELETS AND STATISTICS 55, 59 (Anestis Antoniadis & Georges Oppenheim eds., 1995).

[101] *See, e.g.*, Michèle B. Nuijten, Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp & Jelte M. Wicherts, *The Prevalence of Statistical Reporting Errors in Psychology* (1985–2013), 48 BEHAV. RSCH. METHODS 1205, 1205 (2016) ("One in eight papers contained a grossly inconsistent p-value that may have affected the statistical conclusion.").

[102] Sara Schroter, Nick Black, Stephen Evans, James Carpenter, Fiona Godlee & Richard Smith, *Effects of Training on Quality of Peer Review: Randomised Controlled Trial*, 328 BMJ 673, Table 1 (2004). Nearly all trained subjects completed a second review post-training. The self-taught group caught 3.4 major errors of the nine on average, while the face-to-face trained group caught 3.2 major errors on average. *Id.* at Table 2.

another, by thousands of journals.[103] For instance, many journals now require authors to provide their raw data, analysis code, and other research materials with the submitted manuscript.[104] The guidelines also recommend that journals implement standards for citation to supporting research materials, research design and analysis reporting standards; require public posting of hypotheses and analysis plans prior to observing or collecting data; and encourage submission of studies that attempt to replicate published results.

In addition to disclosing data and other research materials, psychology researchers are increasingly prospectively registering (i.e., preregistering) their studies as a method to avoid questionable research practices.[105] Preregistration entails posting on a public repository the study's protocol, measures, analysis plan, and data exclusion criteria. This is done prior to collecting or observing the data. Preregistration requires a commitment to an analysis plan, making it harder for researchers to selectively report results or selectively exclude data. A second benefit is that other researchers can find studies that have been conducted but not published, possibly due to publication bias. This helps researchers better understand the full scope of what's known and develop a more informed perspective on directions for future work.

The benefits of preregistration have been magnified by editorial boards' recent move to publish Registered Reports.[106] Journals employ a non-traditional peer review process to evaluate Registered Reports. Editors and peer reviewers review the preregistered study before any data are collected or observed. If the "stage one" manuscript is accepted, the editor promises to publish the study as long as the author follows the approved plan. Registered Reports remove the incentive for authors to engage in questionable research practices because they do not have to engage in such practices to produce exciting findings or findings that support their hypotheses. They also eliminate at least some forms of publication bias because the study is published regardless of its results. Over 300 journals now publish Registered Reports, including two legal-psychology journals.[107]

---

[103] Brian A. Nosek et al., *Promoting an Open Research Culture*, 348 SCI. 1422, 1424 (2015); *TOP Guidelines*, CTR. OPEN SCI., https://www.cos.io/initiatives/top-guidelines (last accessed Feb. 7, 2023).

[104] *Id*. Additionally, some student-edited law journals now require research material transparency. *See, e.g.*, *Joint Law Review Statement on Data and Code Transparency*, YALE L. J. https://www.yalelawjournal.org/files/JointLawReviewStatementonDataandCodeTransparency_icaqqmh7.pdf (last accessed Feb. 7, 2023).

[105] Kai Kupferschmidt, *More and More Scientists Are Preregistering Their Studies. Should You?*, SCI. (Sept. 21, 2018), https://www.science.org/content/article/more-and-more-scientists-are-preregistering-their-studies-should-you; Moin Syed, *Three Myths About Open Science That Just Won't Die*, PSYARXIV PREPRINT (2022), https://europepmc.org/article/ppr/ppr573785 (manuscript at 7-9) (arguing preregistration is appropriate for any study design, not just experiments).

[106] *See* Christopher D. Chambers & Loukia Tzavella, *The Past, Present and Future of Registered Reports*, 6 NATURE HUM. BEHAV. 29, 31 (2022) (Reviewing the rise of registered reports from their availability at a few journal to hundreds, and their benefits).

[107] Law and Human Behavior Editorial Team, *Law and Human Behavior: Registered Reports Instructions*, AM. PSYCH. ASS'N, https://www.apa.org/pubs/journals/features/lhb-registered-

Metaresearch studies are finding that preregistration and Registered Reports are working as intended. For instance, Anne Scheel and colleagues recently found that Registered Reports yield a more credible pattern of results than do standard reports.[108] Specifically, they found that while about 95% of studies published using the traditional process supported the author's first hypothesis, the same was the case for only about 40% of Registered Reports. This difference might serve as evidence of researchers' ability to engage in questionable research practices under the traditional publication process and for the existence of considerable publication bias. In addition, some evidence suggests that Registered Reports use stronger methodologies because reviewers can help improve the methods before data are collected. Indeed, an observational study recently found that blind raters judged Registered Reports to be of higher quality across factors such as strength of analysis and rigor and creativity of the methods.[109]

Furthermore, some have found that a combination of preregistration and Registered Reports not only potentially enhances measurement accuracy but also is positively correlated with reports of smaller effect sizes relative to non-preregistered studies. Amanda Kvarven and colleagues compared effect sizes reported in meta-analyses of standard studies to effect sizes found in large registered replication projects (some being Registered Reports).[110] They found that effect measurements for the registered replication projects were one third the size found in the meta-analyses.[111] Schäfer and Schwarz performed a similar analysis, but instead of relying on meta-analyses, they randomly selected 900 psychology studies and compared their effect measurements to all 93 preregistered studies they could find at the time (16 of which were Registered Reports).[112] The

---

reports.pdf (last accessed Feb. 7, 2023); Legal and Criminological Psychology, *Author Guidelines*, WILEY, https://onlinelibrary.wiley.com/page/journal/20448333/homepage/forauthors.html (last accessed Feb. 7, 2023). On the importance of openness and transparency to law and psychology, *see* Bradley McAuliff, Melanie Fessinger, Anthony Perillo & Jennifer Torkildson Perillo, *Psychology and Law, Meet Open Science*, PSYARXIV PREPRINTS (2021), https://psyarxiv.com/8d2tx/ (manuscript at [3-4]).

[108] Scheel et al., *supra* note 66, at 5.

[109] Courtney K. Soderberg, Timothy M. Errington, Sarah R. Schiavone, Julia Bottesini, Felix Singleton Thorn, Simine Vazire, Kevin M. Esterling & Brian A. Nosek, *Initial Evidence of Research Quality of Registered Reports compared with the Standard Publishing Model*, 5 NATURE HUM. BEHAV. 990, 992 (2021).

[110] Kvarven et al., *supra* note 61, at 431-32.

[111] *Id*. at 425. They reported a Cohen's *d*=0.16 for the registered replication projects and a Cohen's *d*=0.42 for the non-preregistered meta-analyses. Cohen's *d* is an effect size measure that gauges the magnitude of differences between groups. See https://osf.io/3zgxe. In addition, in seven of the replications, the effect was no longer statistically significant (whereas it had been in the meta-analysis). Note however that this study is limited due to the small sample. The authors could only find fifteen meta-analyses with matching large replication comparators.

[112] Thomas Schäfer & Marcus A. Schwarz, *The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases*, 10:813 FRONTIERS PSYCH. 1, 4-7 (2019). Data were collected in 2018 and, given the rapid rise of preregistration in social science, we expect that number would be much greater today.

preregistered studies reported effects about 45% as large as the standard studies.[113]

That preregistered studies with larger samples report substantially smaller effects poses a problem for legal actors who rely on findings from psychology studies. Most law and policy decisions involve trade-offs. Reforms are often costly both in terms of resources and lost opportunities to pursue other remedies. While it may be theoretically interesting that some variable has a statistically significant effect on another, whether a particular reform's costs should be incurred often depends on practical significance, which is determined by the measured effect's size.

To further complicate matters, the conventions that psychologists use to categorize effect sizes potentially blur legal and policy implications of reported effects. Psychologists commonly classify effect sizes using cutoffs, referring to them as small (Cohen's $d = 0.2$), medium (Cohen's $d = 0.5$) and large (Cohen's $d = 0.8$).[114] Legal commentators sometimes parrot this language.[115] These arbitrary benchmarks, however, distract from the practical meaning of measured effect sizes in a particular legal and policy setting. A psychological intervention that saves a small number of lives is nevertheless much more important than an intervention that provides a big increase in people's mood. These contextual factors must be taken into account to understand the tradeoffs involved in using and implementing a particular psychology finding.[116]

We now turn from studies that mistakenly conclude that there is an effect, or overestimate the effect size, to studies that mistakenly declare that an effect *does not* exist.

---

[113] They reported a median Pearson's r=0.16 for preregistered studies compared to 0.36 in standard reports. Pearson's r is an effect size measure that gauges the strength of association between variables. Pearson's r ranges from negative one (a perfect negative correlation) to zero (no relationship) to one (a perfect positive correlation). See https://osf.io/3zgxe.

[114] Daniël Lakens, *Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: A Practical Primer for T-Tests and ANOVAs*, 4 FRONTIERS PSYCH. 863, 3 (2013).

[115] *See, e.g.*, Jerry Kang, *What Judges Can Do about Implicit Bias*, 57 CT. REV. 78, 80 (2021) ("...IAT scores predict intergroup discriminatory behavior at a very low level. (…By convention, r values greater than or equal to 0.1, 0.3, and 0.5 are called small, medium, and large, respectively) The small effect size that has been found should not be surprising…"); Masaki Iwasaki, *Relative Impacts of Monetary and Non-Monetary Factors on Whistleblowing Intention: The Case of Securities Fraud*, 22 U. PA. J. BUS. L. 591, 613 (2020) (recognizing that "[s]tandardized coefficients are used to compare the *relative* impacts of independent variables" [emphasis added], but going on to describe effect sizes in absolute terms: "Conventionally, the standardized coefficients of 0.1, 0.3, and 0.5 represent small, medium, and large effect sizes respectively: a medium effect size is "likely to be apparent to the naked eye of a careful observer"; a small effect size is "noticeably smaller yet not trivial"; a large effect size is "the same distance above medium as small is below it." (citing Jacob Cohen, *Statistical Power Analysis*, 1 CURRENT DIRECTIONS PSYCH. SCI. 98, 99 (1992)); Andrew Jurs, *Gatekeeper with a Gavel: A Survey Evaluating Judicial Management of Challenges to Expert Reliability and Their Relationship to Summary Judgement*, 83 MISS. L.J. 325, 363 (2014) ("These r-values indicate a small-to-medium effect size under Cohen's categorical definitions").

[116] *See* Jason M. Chin, Effect Sizes, *Law and Psychology Must Think Critically About Effect Sizes*, 3 DISCOVER PSYCH. 1 (2023), https://link.springer.com/article/10.1007/s44202-022-00062-2..

B. What to make of null findings?

Credible null findings–findings that a particular manipulation has no effect or, more generally, that a particular claim is not true–are useful in both psychology and in law. For instance, a recent study found that having expert witnesses give evidence concurrently did *not* mitigate the well-documented[117] problem of experts providing biased testimony due to adversarial allegiance.[118] Similarly, a large government-funded study in Australia found that, in sexual assault cases with multiple complainants, joining trials such that those complainants gave evidence in the same proceeding did *not* unduly bias (mock) jurors against the accused.[119] While null findings like these are difficult to publish under traditional publication models given editor bias in favor of statistically significant findings, they might assist courts and lawmakers by providing some evidence that a particular reform is not worth the effort (as with concurrent experts) or by demonstrating that a more efficient court procedure does not present a downside that some argue it might (as with joining trials). Studies that report null results, however, are useful only if they were designed such that they *could* have found a meaningful effect if, in fact, one exists. To return to our telescope analogy, pointing a telescope at the sky and seeing nothing does not necessarily justify the claim that there really is nothing there if the telescope has smudges all over the lens.

Insufficient sample sizes are a predominant driver behind false negative findings. Given the importance of drawing sufficiently large samples, metaresearchers have exhorted psychologists to carefully plan and justify their sample sizes.[120] For applied work, including psychology applied to law, researchers arguably should plan their sample sizes based on the smallest effect size of interest.[121] Studies that include enough participants to find the smallest effect size of interest make null findings meaningful by allowing researchers to confidently conclude that an effect large enough to matter in practice does not exist. For instance, legal practitioners might generally agree that misremembering three details constitutes cause to challenge a witness's credibility in court. Accordingly,

---

[117] *See generally*, National Justice Compania Naviera SA v Prudential Assurance, [1993] F.S.R 563, 565; Jason M Chin, Michael Lutsky & Itiel E Dror, *The Biases of Experts: An Empirical Analysis of Expert Witness Challenges*, 42(4) MAN. L.J. 21, 30 (2019) (detailing the history of judicial concerns with expert witness biases).

[118] Jennifer T. Perillo, Anthony D. Perillo, Nikoleta M. Despodova & Margaret Bull Kovera, *Testing the Waters: An Investigation of the Impact of Hot Tubbing on Experts from Referral Through Testimony*, 45 LAW & HUM. BEHAV. 229, 229 ("Concurrent testimony did not eliminate adversarial allegiance") (2021).

[119] For a review of this study, *see* Jill Hunter & Richard I. Kemp, *Proposed Changes to the Tendency Rule: A Note of Caution*, 41 CRIM. L.J. 253, 257-59 (2017).

[120] *See* Daniel Lakens, *Sample Size Justification*, 8(1) COLLABRA: PSYCH. 1, 1 (2022).

[121] In psychology generally, *see* Carmel Camilleri, Nataly Beribisky, & Rob Cribbie, *The Minimally Meaningful Effect Size: A Vital Component of Pre-Registrations*, PSYARXIV PREPRINTS (2022), https://psyarxiv.com/jbgtm/. In legal psychology, *see* Henry Otgaar, Paul Riesthuis, Johannes G. Ramaekers, Maryanne Garry & Lilian Kloft, *The Importance of the Smallest Effect Size of Interest in Expert Witness Testimony on Alcohol and Memory*, 13 FRONTIERS PSYCH. 1, 4 (2022) ("One way to accomplish these aims is to decide on the smallest effect size of interest").

researchers testing interview techniques to determine if they produce misremembered details of an event should include enough participants to be sufficiently powered to detect an effect size of three misremembered details. If the study fails to find an effect, researchers can convincingly conclude that the technique does not increase misremembrances to a meaningful degree.

Unfortunately for legal actors, psychology researchers seem to rarely justify their sample sizes based on the smallest effect size of interest.[122] Rather, researchers revert to rules of thumb for planning their sample sizes, such as including 15-20 participants per condition regardless of what they are studying.[123] Some researchers do calculate how many participants are needed to find a "medium" effect size,[124] but as we saw above, effect size labels are unrelated to practical effect sizes. Sample size justification based on the smallest effect of interest is a better practice and is indeed an established practice in other applied fields.[125]

While psychology seems to be moving in the right direction by thinking more critically about sample sizes, much more work is required. In legal psychology, a group of false memory researchers recently set out to systematically study what their field considered to be the smallest effect size of interest to guide future sample size planning.[126] They asked researchers, for instance, about what effect size from classic studies in the field would be the smallest but still theoretically and practically meaningful. They found no evidence for consensus in the field. More troublingly, for some prompts they found that the plurality of respondents said that the smallest effect size of interest is any result that is statistically significant.[127] As the authors of the study note, this result indicates that many false memory researchers are not well-trained in study design: "based on our findings, memory researchers seemed to conflate statistical significance (i.e., $p < .05$) with a practically meaningful effect size. This is concerning because statistical significance is not the same as a practically meaningful effect."[128]

---

[122] Marjan Bakker, Chris H. J. Hartgerink, Jelte M. Wicherts, & Han L. J. van der Maas, *Researchers' Intuitions About Power in Psychological Research*, 27(8) PSYCH. SCI. 1069, 1074 (2016) ("When asked about how they normally determined sample sizes in their own studies, more than half of our respondents indicated that they did not use a power analysis, which may explain why such analyses are presented in fewer than 3% of psychological articles.").

[123] *Id*. at 1074; Marjan Bakker, Coosje L. S. Veldkamp, Olmo R. van den Akker, Marcel A. L. M. van Assen, Elise Crompvoets, How Hwee Ong, & Jelte M. Wicherts, *Recommendations in Pre-Registrations and Internal Review Board Proposals Promote Formal Power Analyses but Do Not Increase Sample Size*, 15 PLOS ONE 1, 7 (2020).

[124] Bakker et al., Veldkamp, *supra* note 123, at 11-12.

[125] *See generally,* Madeleine T. King, *A Point of Minimal Important Difference (MID): A Critique of Terminology and Methods,* 11 EXPERT REV. OF PHARMACOECONOMICS & OUTCOMES RSCH. 171, 172-174 (2011); Richard S. E. Keefe, Helena C. Kraemer, Robert S. Epstein, Ellen Frank, Ginger Haynes, Thomas P. Laughren, James McNulty, Shelby D. Reed, Juan Sanchez & Andrew C. Leon, *Defining a Clinically Meaningful Effect for the Design and Interpretation of Randomized Controlled Trials*, 10 INNOVATIONS CLINICAL NEUROSCIENCE 4s, 18s (2013).

[126] Paul Riesthuis, Ivan Mangiulli, Nick Broers & Henry Otgaar, *Expert Opinions on the Smallest Effect Size of Interest in False Memory Research*, 36 APPLIED COGNITIVE PSYCH. 203, 203 (2022).

[127] *Id*. at 208.

[128] *Id*. at 212.

In addition to the uncertainty about the existence and size of experimental psychology findings, the current sample size customs in experimental psychology further limit the field's applicability to law. Insufficient samples that produce false negatives cause at least two problems that legal actors must keep in mind. First, some of these findings end up in the file drawer because editors are reluctant to publish null results. In the case of insufficient samples, this is the ideal publication outcome, but to the extent that these flawed studies end up in a public file drawer (a repository that collects preregistered studies and their results), researchers who rely on the public file drawer to get a complete understanding of the existing literature might misinterpret the null result as valid or might feel compelled to perform the study again using a sufficient sample. The latter highlights the waste of time and funds expended on the original faulty study, which delays the already-slow scientific process required to understand how best to develop law and policy. Second, in rare cases, the false negative findings might find their way into publication. Here, those who wish to import the findings into law must be able to distinguish between true null results and false negatives.

C. Questionable measurement practices

Psychologists have invented ways of measuring and quantifying psychological processes that are impossible to directly observe.[129] Indeed, the law and psychology literature contains many scales and tools for measuring a variety of psychological processes, many of them highly relevant to the legal system. These include psychopathy,[130] well-being,[131] and attitudes towards marginalized groups.[132] Measurement scales and tools can be useful to both practitioners and researchers. For instance, checklists used for clinical diagnoses can help ensure more reliable diagnoses.

As we saw in the case of research design and analysis, researchers can, consciously or subconsciously, leverage undisclosed flexibility to produce a desired result. The same is true with measurement. Specifically, questionable *measurement* practices are "decisions researchers make that raise doubts about the validity of the measures used in a study, and ultimately the validity of the final conclusion."[133] Researchers make many such decisions related to measurement, and these can limit our ability to usefully apply eventual findings to law. Jessica Kay Flake and Eiko Fried note, for instance, that "data collected with a 10-item questionnaire can be

---

[129] *See* Tess M. S. Neal, Kristy A. Martire, Jennifer L. Johan, Elizabeth M. Mathers & Randy K. Otto, *The Law Meets Psychological Expertise*: *Eight Best Practices to Improve Forensic Psychological Assessment*, 18 ANN. REV. L. & SOC. SCI. 169, 170-171 (2022).
[130] Neal et al., *supra* note 4, at 149.
[131] *See* Lawrence S. Krieger & Kennon M. Sheldon, *What Makes Lawyers Happy: A Data-Driven Prescription to Redefine Professional Success*, 83 GEO. WASH. L. REV. 554, 562 (2014) ("We employed the concept of 'subjective wellbeing' ('SWB') to measure happiness in this study").
[132] Kang et al., *supra* note 2, at 1130.
[133] Jessica Kay Flake & Eiko I. Fried, *Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them*, 3(4) ADVANCES METHODS & PRAC. PSYCH. SCI. 456, 458 (2020).

summarized in one sum score 1,023 different ways by summing different subsets of items (e.g., using Items 1–3 to calculate a sum score, or using Items 2–10 or Items 6–8, and so on), and even more possibilities exist if multiple subscales or different analytic techniques can be used to obtain a final score."[134] To modify our running analogy, whereas using a low-powered study is akin to using a telescope with a smudge on the lens, using inappropriate measurement practices is akin to using a microscope to look at the sky rather than a telescope.

Unfortunately, there is ample evidence for questionable measurement practices in both experimental psychology and law and psychology. Flake and Fried reviewed several questionable measurement practices used in experimental psychology, finding that they are "ubiquitous, are largely ignored in the literature, provide researchers with ample degrees of freedom that can be exploited to obtain desired results, and thus pose a serious threat to cumulative psychological science."[135] For example, Amy Orben and Andrew Przybylski conducted a review of the research linking digital device use to adolescent well-being.[136] In measuring well-being, they found that researchers engaged in a great deal of cherry picking between and within measures such that the decisions produced "many different possibilities for combining and analysing these measures, making the pre-specified constructs more of an accessory for publication than a guide for analyses."[137]

Within legal psychology, Tess Neal and colleagues have raised several issues with questionable measurement practices and courts' failure to police them.[138] This state of affairs is worrisome because measurement plays an important role in legal psychology, with assessments contributing to important decisions, such as parental fitness,[139] defenses based on mental illness,[140] and the risk a person poses to society.[141] While some assessment tools have been tested and shown to lead to consistent outcomes across practitioners, psychologists' choices about which tool to use "vary widely."[142] In addition, only about 60% of tools currently in wide use have been exposed to robust peer review.[143] Despite this, parties rarely challenge the admissibility of psychological assessments, and when they do, they are rarely excluded.[144] For instance, U.S. judges still regularly admit into

---

[134] *Id*.
[135] *Id*. at 457.
[136] Amy Orben & Andrew K. Przybylski, *The Association Between Adolescent Well-Being and Digital Technology Use*, 3 NATURE HUM. BEHAV. 173, 173 (2019).
[137] *Id*. at 181 (analyzing how twelve studies used measures from a large dataset including measures of technology use and adolescent well-being).
[138] Neal et al., *supra note* 4, at 135 ("Challenges to the most scientifically suspect tools are almost nonexistent. Attorneys rarely challenge psychological expert assessment evidence, and when they do, judges often fail to exercise the scrutiny required by law.").
[139] Lefkowitz v. Ackerman, No. 2:16-cv-00624, 2017 WL 4237068 (S.D. Ohio Sept. 25, 2017).
[140] People v. Jing Hua Wu, No. H040066, 2016 WL 616744 (Cal. Ct. App. Feb. 16, 2016).
[141] Wisconsin v. Loomis, 881 N.W.2d 749 (Wis. 2016).
[142] Neal et al., *supra* note 4, at 147.
[143] Neal et al., *supra* note 4, at 143.
[144] *Id*. at 153.

evidence results of the Rorschach Inkblot Test despite substantial scientific critiques calling the test's validity into question.[145]

In addition, in most cases, assessment tools, and measures more generally, have yet to be tested in contexts outside the lab or in situations that closely resemble their use outside the lab.[146] We turn to this fourth limitation next.

D. Limited and unknown generalizability

According to one view of the research-to-action pipeline, psychological phenomena should first be studied in the controlled confines of psychology laboratories, or nowadays, in participants' homes as they engage in studies presented through various online platforms.[147] Laboratory study findings should then be replicated in the lab using different materials and formats to determine whether the original findings were artefacts of a particular research setup. Researchers should then study the same phenomena in the field to determine if they generalize to scenarios closely resembling the context of interest. Unfortunately, as we will now discuss, this progression is rare, with, for the most part, only inchoate attempts at providing safe generalizations or an understanding of why psychology findings may be specific to particular contexts and populations.[148] Several studies have found that when social science is taken outside of the lab, observed effects are smaller or non-existent.[149]

Although psychology theories and claims are often broad, covering a variety of contexts and tasks, they are tested by measuring quantitative performance on specific tasks using specific stimuli.[150] For example, legal psychologists have long been interested in a finding known as "verbal

---

[145] *Id*. at 149.

[146] Neal et al., *supra* note 129 at 177 ("Overall, little empirical evidence is available regarding the field validity of the methods used by practitioners to produce forensic psychological opinions") *See generally* John F. Edens & Marcus T. Boccaccini, *Taking Forensic Mental Health Assessment "Out of the Lab" and Into "The Real World": Introduction to the Special Issue on the Field Utility of Forensic Assessment Instruments and Procedures*, 29 PSYCH. ASSESSMENT 599, 600 (2017) (bemoaning the dearth of studies testing the validity of psychological measures outside the lab).

[147] Christopher J. Bryan, Elizabeth Tipton & David S. Yeager, *Behavioural Science is Unlikely to Change the World Without a Heterogeneity Revolution*, 5 NATURE HUM. BEHAV. 980, 980-981 (2021); Thomas V. Pollet & Tamsin K. Saxton, *How Diverse Are the Samples Used in the Journals 'Evolution & Human Behavior' and 'Evolutionary Psychology'?*, 5 EVOLUTIONARY PSYCH. SCI. 357, 362 (2019) ("we need to compare individuals from different ecological settings").

[148] *See generally,* Bryan et al., *supra* note 147, at 980.

[149] *A Confirmation Prompt Reduced Financial Self-Reporting Errors Initially, but the Effect Did Not Persist in Subsequent Periods, Office of Evaluation Sciences*, OFF. EVALUATION SCIS., https://oes.gsa.gov/projects/iff-confirmation-prompt-update/ (last updated Dec. 1, 2021); Alan Gerber, Gregory Huber & Albert Fang, *Do Subtle Linguistic Interventions Priming a Social Identity as a Voter Have Outsized Effects on Voter Turnout? Evidence from a New Replication Experiment*, 39 POL. PSYCH. 925, 934 (2018) ("the turnout levels in the noun and verb wording groups are statistically indistinguishable"); Hunt Allcott, *Site Selection Bias in Program Evaluation*, 130 Q. J. ECON. 1117, 1117 (2015). Note that we do not know how representative these examples are.

[150] Yarkoni, *supra* note 30, at 1 (noting that most psychological theories are broad and general and cannot easily be tested using specific research setups).

overshadowing,"[151] which is the impairment of facial recognition that we seem to experience after we describe an observed face.[152] This effect is counterintuitive, because one might instead expect that reciting one's memory of a face would help cement that memory. Verbal overshadowing researchers, on the other hand, posited that because faces are difficult to describe in words, verbalization might cause people to generate verbal descriptions that do not match their original experience. These faulty descriptions can taint the witness's responses to interview questions related to identification. This could have practical consequences for the use of and interpretation of police interviews.

While verbal overshadowing initially found experimental support, several researchers reported difficulty replicating the effect, and subsequently a large multi-lab collaboration of researchers investigated it. In one of the first efforts in psychology to use the Registered Report format, the research team found that while verbal overshadowing did replicate, the effect was undetectable when there was a 20-minute gap between verbalization and selecting the target from the lineup.[153] As Tal Yarkoni put it, one specific research setup "cannot provide a meaningful test of a broad construct like verbal overshadowing."[154]

Verbal overshadowing's lack of robustness would not be as problematic were it regular practice to systematically test theories to assess their generalizability. Many psychology theories and claims, however, have been tested using a small set of designs and a limited set of stimuli. As a result, psychologists simply cannot say whether those theories and claims have been exposed to testing that might have falsified them. Yarkoni refers to this limitation as psychology's "generalizability crisis."[155] :

Limits on generalizability also flow from a mismatch between laboratory conditions and conditions outside the lab.[156] Consider the many assessment tools, such as checklists, used by psychologists and psychiatrists in their work in the legal system. Clinicians use these tools to diagnose legally relevant conditions to assist factfinders on issues including the risk posed by those in the criminal justice system and whether an accused is fit

---

[151] Jason M. Chin & Jonathan W. Schooler, *Why Do Words Hurt? Content, Process, and Criterion Shift Accounts of Verbal Overshadowing*, 20 EUROPEAN J. COGNITIVE PSYCH. 396, 396 (2008).
[152] Stimulus variability is also an issue in implicit bias research, as we discuss *infra* Part II. *See* Charles M. Judd, Jacob Westfall & David A. Kenny, *Treating Stimuli As a Random Factor in Social Psychology: A New and Comprehensive Solution to a Pervasive but Largely Ignored Problem*, 103 J. PERSONALITY & SOC. PSYCH. 54, 54-57 (2012).
[153] V. K. Alogna et al., *Registered Replication Report: Schooler and Engstler-Schooler* (1990), 9 PERSP. PSYCH. SCI. 556, 566 (2014).
[154] Yarkoni, *supra* note 30, at 8. For questions about whether behavioral economics can be generalized to legal problems, *see* Robert A. Hillman, *Limits of Behavioral Decision Theory in Legal Analysis: The Case of Liquidated Damages*, 85 CORNELL L. REV. 717, 718 (2000).
[155] Yarkoni, *supra* note 30, at 1.
[156] Simine Vazire, Sarah R. Schiavone & Julia G. Bottesini, *Credibility Beyond Replicability: Improving the Four Validities in Psychological Science*, 31(2) CURRENT DIRECTIONS PSYCH. SCI. 162, 165 (2022) (reviewing the many reasons laboratory findings may not generalize.); Neal et al., *supra* note 4, at 177; Gregory Mitchell, *Revisiting Truth or Triviality: The External Validity of Research in the Psychological Laboratory*, 7 PERSPS. PSYCH. SCI. 109, 109 (2012).

to stand trial. While many of these tools have been shown to work well[157] in the laboratory, their "field validity"–"whether a tool or method is accurate, repeatable, and reproducible under routine practice conditions typical of real-world work"–remains an open question.[158]

Field validity might lag laboratory results for several reasons.[159] First, initial laboratory studies are often designed to demonstrate that a tool has desirable properties. For this reason, the researchers might choose only highly trained assessors or exclude assessors who perform poorly in initial tests.[160] Second, adversarial pressures to come to a certain conclusion, such as pressure that retaining lawyers exert on retained psychologists, can affect the results.[161] Third, individuals being assessed in the field may also be more circumspect or engage in impression management rather than being entirely truthful when the assessment poses serious consequences.[162]

Although field conditions likely decrease the reliability of psychological assessment tools, very little is known whether and to what degree this occurs because field studies are rare.[163] What little research exists suggests that assessments conducted in the field are considerably less reliable than those conducted in the lab.[164] For instance, field studies find that the widely-used assessment tool Psychopathy Checklist-Revised is more unreliable in the field relative both to the laboratory and to the reliability estimate reported in its user's manual.[165]

Clinical assessment research is not unique in this respect. Ample evidence suggests that other laboratory research also does not necessarily replicate reliably in the field.[166] For instance, Gregory Mitchell identified 82 meta-analyses of psychology studies that include both lab and field

---

[157] By work well, we mean that, for instance, multiple assessors using the same tool in the lab come to similar conclusions.
[158] Neal et al., *supra* note 129, at 173.
[159] Edens & Boccaccini, *supra* note 146, at 601-02.
[160] *Id*. at 601.
[161] Daniel C. Murrie, Marcus T. Boccaccini, Lucy A. Guarnera & Katrina A. Rufino, *Are Forensic Experts Biased by the Side that Retained Them?*, 24 Psych. Sci. 1889, 1895 (2013).
[162] Edens & Boccaccini, *supra* note 146, at 602.
[163] Lucy A. Guarnera & Daniel C. Murrie, *Field Reliability of Competency and Sanity Opinions: A Systematic Review and Meta-Analysis*, 29 PSYCH. ASSESSMENT 795, 809 (2017); Edens & Boccaccini, *supra* note 146, at 602 ("one might reasonably ask: why are there so few of them published in mainstream research outlets").
[164] Assessments are deemed to be less reliable in the field if multiple assessors of the same individual tend to agree less often than they do in the lab. W. Neil Gowensmith, Daniel C. Murrie, Marcus T. Boccaccini & Brandon J. McNichols, *Field Reliability Influences Field Validity: Risk Assessments of Individuals Found Not Guilty by Reason of Insanity*, 29 PSYCH. ASSESSMENT 786, 790 (2017) ("This reflects a 'poor' level of agreement"); W. Neil Gowensmith, Stephanie N. Sessarego, Meghan K. McKee, Samantha Horkott, Nina MacLean & Katherine E. McCallum, *Diagnostic Field Reliability in Forensic Mental Health Evaluations*, 29 PSYCH. ASSESSMENT 692, 696 (2017) ("…evaluators agreed on a defendant's entire diagnostic picture in fewer than one of five cases.").
[165] John F. Edens, Jennifer Cox, Shannon Toney Smith, David DeMatteo, & Karolina Sörman, *How Reliable Are Psychopathy Checklist–Revised Scores in Canadian Criminal Trials? A Case Law Review*, 27 PSYCH. ASSESSMENT 447, 452, 452 (2015) ("reliability of the PCL-R in Canadian criminal cases is considerably lower than what is reported in the instrument's professional manual.").
[166] *See* Mitchell, *supra* note 156, at 112.

studies. He found a fairly strong correlation between the results of lab and field studies in some subfields in psychology but not in others.[167] Specifically, he found a lab-field correlation of $r = 0.89$ for industrial organizational psychology. This subfield of psychology studies how psychological processes operate in the workforce and has a long tradition of testing robustness in the field.[168] Mitchell reported a weaker correlation ($r = 0.53$) for social psychology research–the subfield that legal scholars often point to for evidence supporting many legally relevant phenomena, such as implicit bias.[169] Perhaps most surprising, for 29% of the findings in social psychology, Mitchell found a sign reversal. In other words, positive relationships between stimulus and response in the lab reversed direction in the field (or vice versa), suggesting that use in the field is actually *counterproductive*.

Finally, metaresearch is uncovering a discouraging lack of diversity among researchers and research participants.[170] This further indicates limited generalizability to people beyond those typically studied in psychology research. For example, in a study of psychology articles from leading journals published between 2014-2018, 81% of first authors and 79% of participants were from the U.S. and other English-speaking countries.[171]

Law and psychology is no exception to the problem of overly narrow sampling. Several researchers who study the psychology of investigative interviewing recently noted that "**almost without exception** the behaviours used to build and measure rapport [in investigative interviewing] draw on theories developed with reference to interpersonal interactions and communication styles in Western contexts or using data drawn from Western, educated, industrialized, rich, and democratic (WEIRD) samples."[172] And, as we discuss in Part II, studies testing ways to reduce discrimination and implicit bias have been overly focused on university students, to their detriment.[173]

In addition to nationality, other sample characteristics, such as age and socio-economic status, are also important for understanding

---

[167] *Id*. at 112.
[168] *Id*. at 115.
[169] We turn to the implicit bias literature in more detail *infra* Part II.
[170] *See* Henrich et al., *supra* note 22, at 63; Amber Gayle Thalmayer, Cecilia Toscanelli & Jeffrey Arnett, *The Neglected 95% Revisited: Is American Psychology Becoming Less American?*, 76 AM. PSYCH. 116, 116 (2021); Hans IJzerman, Natalia Dutra, Miguel Silan, Adeyemi Adetula, Dana M. Basnight Brown, & Patrick Forscher, *Psychological Science Needs the Entire Globe, Part 1: The Problem With U.S. Dominance in Psychological Science*, ASS'N PSYCH. SCI. (Aug. 30, 2021), https://www.psychologicalscience.org/observer/global-psych-science; Elizabeth Levy Paluck & Donald P. Green, *Prejudice Reduction: What Works? A Review and Assessment of Research and Practice*, 60 ANN. REV. PSYCH. 339, 350 (2009).
[171] Thalmayer et al., *supra* note 170, at 31-32.
[172] Lorraine Hope et al., *Urgent Issues and Prospects at the Intersection of Culture, Memory, and Witness Interviews: Exploring the Challenges for Research and Practice*, 27 LEGAL & CRIMINOLOGICAL PSYCH. 1, 9 (2022) [**emphasis in original**].
[173] Paluck & Green, *supra* note 170, at 350 (reviewing research findings that university students exhibit less prejudice than the general population, which makes them dubious targets of study for research with applications outside the lab).

investigated effects. As mentioned *supra* Part I.A., publication bias has contributed to the overestimation of the effects of nudging and choice architecture, highlighting the need for preregistration (to mitigate questionable research practices that are more likely to result in spurious statistically significant results) and Registered Reports (to mitigate editor bias based on results). Yeager and colleagues have illuminated a second limitation of this literature. They reported evidence that suggests that nudge effects are heterogeneous, with college-aged, higher-income samples producing stronger effects.[174] Despite this, the authors of nudge studies have failed to adequately describe their samples, with just 18% of studies providing "even minimal information about characteristics that might moderate effects."[175]

More broadly, reviews of promising intervention studies (e.g., mobile-phone-based reminders, microfinancing) in psychology[176] find that these studies often do not report enough information to allow policymakers to implement them (e.g., missing information about key features of the intervention administering organization such that users are unable to assess whether their own organizations are sufficiently analogous). Unfortunately, this persists despite many real-world implementations of social science theories showing little or no effect despite promising laboratory findings.[177] As Premachandra and Lewis note, incomplete reporting is wasteful: "So although it is wonderful for psychologists to develop and test interventions and publish them in peer-reviewed journals, if we do not share the information that is required by implementers, **the immense potential our work has for fostering change will not be realized**."[178]

E. The nondiagnosticity of "generally accepted" psychology findings

Historically, the primary criterion for admissibility of expert evidence (including psychology expertise) into federal courts in the U.S. was "general acceptance in the particular field in which it belongs."[179] General acceptance remains an important factor both in the United States

---

[174] David S. Yeager, Jon A. Krosnick, Penny S. Visser, Allyson L. Holbrook & Alex M. Tahk, *Moderation of Classic Social Psychological Effects by Demographics in the U.S. Adult Population: New Opportunities for Theoretical Advancement*, 117 J. PERSONALITY & SOC. PSYCH. e84, e91-e96 (2019). Chris Brewin has argued that psychology research cited by expert witnesses in courtrooms is too general to support their opinions. Chris R. Brewin, *Impact on the Legal System of the Generalizability Crisis in Psychology*, 45 BEHAV. & BRAIN SCIS. e7, 23-24 (2022) ("The generalizability crisis […] can lead to particularly unfortunate consequences in applied fields such as the law.").

[175] Bryan et al., *supra* note 147, at 981.

[176] Bharathy Premachandra & Neil A. Lewis Jr., *Do We Report the Information That Is Necessary to Give Psychology Away? A Scoping Review of the Psychological Intervention Literature 2000–2018*, 17 PERSP. PSYCH. SCI. 226, 232 (2022) (noting, for instance, that only thirteen percent report how much intervention costs).

[177] Bryan et al., *supra* note 147, at 980 n.24-29.

[178] Premachandra & Lewis, *supra* note 176, at 234 (emphasis added).

[179] Frye v. United States, 293 F. 1013 (D.C. Cir. 1923).

and abroad.[180] Perhaps not surprisingly then, Wigmore, as an evidence scholar, focused much of his critique on whether Münsterberg's views were as generally accepted as he claimed.[181]

To assess general acceptance, legal psychologists conduct periodic audits to determine what percentage of researchers agree with a particular claim or finding.[182] While these studies are interesting because they help us understand what the field believes at a given time, the lessons from psychology's replication crisis suggest that audits should be viewed with caution (a caution the audit authors do not regularly provide).[183] One critical lesson from recent replication projects is that, commonly, findings that have reached general acceptance cannot be replicated, even those findings that seem robust in meta-analyses.[184]

F. Scientific communication in psychology

> Together, these issues create an interesting dilemma for science communication in psychology. **If we cannot consistently reproduce a large portion of our own findings, and it is unclear the extent to which those findings generalize beyond the narrow slivers of the population** (Syed & Katha-walla, 2020) **and toy problems that we often study** (Dunnette, 1966; Navarro, 2019), then what exactly are we to "give away?" **What should we be communicating to the broader public**? And precisely how are we to do that?[185]

Many of the limits we have discussed likely are not obvious to a non-expert when reading psychology studies, systematic reviews, and popular psychology books. Some of these limits arise from subtle methodological choices that many psychologists themselves did not realize

---

[180] The case that overruled Frye, Daubert v. Merrell Dow Pharms., 509 U.S. 579 (1993), retained general acceptance as a factor for determining the reliability of expert evidence. Canadian courts regularly apply Daubert and its general acceptance factor. *See* R. v. Trochym, 2007 SCC 6.
[181] Wigmore, *supra* note 2, at 412-16.
[182] Saul M. Kassin, Allison D. Redlich, Fabiana Alceste & Timothy J. Luke, *On the General Acceptance of Confessions Research: Opinions of the Scientific Community*, 73 AM. PSYCH. 63, 63 (2018); Saul M. Kassin, V. Anne Tubb, Harmon M. Hosch & Amina Memon, *On the "General Acceptance" of Eyewitness Testimony Research: A New Survey of the Experts*, 56 AM. PSYCH. 405, 405 (2001).
[183] The authors of the general acceptance surveys do not caution about publication bias, questionable research practices, or other credibility-related factors that may limit the usefulness of general acceptance.
[184] *See* Kathleen D. Vohs et al., *A Multisite Preregistered Paradigmatic Test of the Ego-Depletion Effect*, 32(10) PSYCH. SCI. 1566, 1574 (2021) (reporting no statistically significant effect); Katie E. Garrison, David Tang & Brandon J. Schmeichel, *Embodying Power: A Preregistered Replication and Extension of the Power Pose Effect*, 7 SOC. PSYCH. & PERSONALITY SCI. 623, 625-26 (2016). For a systematic comparison of meta-analyses to large replication studies, *see* Kvarven et al., *supra* note 61, at 425-429.
[185] Neil A. Lewis Jr. & Jonathan Wai, *Communicating What We Know and What Isn't So: Science Communication in Psychology*, 16 PERSP. PSYCH. SCI. 1242, 1243 (2021) (emphasis added).

were problematic until recently.[186] Others, such as limited generalizability, are difficult to detect because authors "rarely identify the target populations for their inferences" and "fewer justify their (often implicit) claims of generality."[187] One way that at least some academic work makes these uncertainties more salient is through explicit statements of limitations.

Researchers have assessed how often limitations are noted in psychology studies. A recent study by Clarke and colleagues examined all psychology studies published from 2010-2020 in a high impact, short format journal (*Social Psychological and Personality Science*) and found that the average article described few limitations (1.5 per article).[188] They found that the type of limitations mentioned varied. One in four suggested the study may not generalize to other contexts and situations. Other warnings were rarer, including limitations related to sample size (1 in 15 articles) and limitations related to effect size (1 in 20).[189] Arguably, these warnings should appear in every study.

Psychologists are also now debating another aspect of scientific communication: whose voices should be heard. In particular, researchers have raised concerns about the outsized role that eminent psychologists have in engaging with the public: "eminence does not appear to be a predictor of producing more credible (e.g., replicable) research [...] What we have learned from the meta-science movement in the field over the past decade is that the most efficient strategies for optimizing the metrics that lead to fame are strategies that run counter to the most efficient strategies for producing research that is credible, generalizable, and useful outside of our disciplinary bubble."[190] Accordingly, courts and policymakers risk placing oversized weight on research of high-status psychologists.

In sum, given recent discoveries by metaresearchers that have illuminated the questionable credibility of experimental psychology findings, normative scholars and policymakers must take care when applying such findings to solve problems in law. To crystalize this general advice, Part II analyzes the specific case of how findings from implicit bias studies have been applied in law.

II. Implicit bias in law, a case study

> Using experimental methods in laboratory and field studies, researchers have provided convincing evidence that implicit biases **exist**, are **pervasive**, are **large in magnitude**, and **have real-world effects**. These fascinating discoveries, which have **migrated from the science journals into the law reviews** and even popular discourse, are now **reshaping**

[186] *See* John et al., *supra* note 78, at 524.
[187] Simons et al., *supra* note 43, at 1123.
[188] Beth Clarke, Sarah R. Schiavone & Simine Vazire, *What Limitations Are Reported in Short Articles in Social and Personality Psychology?,* PSYARXIV PREPRINTS (July 1, 2022), https://psyarxiv.com/n4eq7/ (manuscript at 41).
[189] Clarke et al., *supra* note 188, at 46-49.
[190] Lewis & Wai, *supra* note 185, at 1244.

> **the law's fundamental understandings of discrimination and fairness**.[191]
>
> Although critics have expressed concerns about the meaning and significance of the implicit-bias construct for more than a decade (e.g., Arkes & Tetlock, 2004; Fiedler, Messner, & Bluemke, 2006), skeptical views have received significantly more attention over the past few years. In fact, the growing skepticism has become so pervasive that even early proponents have started to question the explanatory value of implicit bias (e.g., Forscher, Mitamura, Dix, Cox, & Devine, 2017), with some critics dismissing the construct as entirely irrelevant for the psychological understanding of social discrimination (e.g., Blanton & Jaccard, 2017; G. Mitchell, 2018).[192]

Implicit bias refers to "automatically activated associations about social groups" [193] and, as the first epigraph above suggests, has potentially serious implications for law. Contrast that, however, with the second epigraph, published just seven years later but providing a remarkably different assessment of implicit bias and its implications for law and society. The first, taken from a widely-cited 2012 law review article,[194] confidently states that implicit bias effects are pervasive and have large, real world effects. The rest of that article makes few attempts to note limitations of those claims or discuss any uncertainty in the research.[195] The second epigraph, on the other hand, alludes to both longstanding and more recent uncertainty about whether implicit bias can help explain social discrimination at all.

---

[191] Kang et al., *supra* note 2, at 1126.

[192] Gawronski, *supra* note 34, at 574; *see also* Gregory Mitchell, *An Implicit Bias Primer*, 25 VA. J. SOC. POL'Y & L. 27, 33-35 (2018) (summarizing conceptual problems with implicit bias and its application to law).

[193] Forscher & Devine, *supra* note 33, at 1.

[194] It has been cited 841 times as of this writing. *Implicit Bias in the Courtroom*, GOOGLE SCHOLAR, https://scholar.google.com.au/scholar?cites=6227791353308058574&as_sdt=2005&sciodt=0,5&hl=en (last accessed Jan. 12, 2023).

[195] The cautions Kang et al., *supra* note 2 provide do not address biases in the literature affecting the size and existence of effects they review (e.g., publication bias) and are typically minimized shortly after the caution is given. We assembled the cautions they do provide, see https://osf.io/zuy8m. Although we focus here on discoveries calling implicit bias into question following the publication of Kang et al. in 2012, strong critiques of the science were published as early as 2009. *See e.g.*, Hart Blanton, James Jaccard, Jonathan Klick, Barabara Mellers & Gregory Mitchell, *Strong Claims and Weak Evidence: Reassessing the Predictive Validity of the IAT*, 94 J. APPLIED PSYCH. 567, 569-578 (2009) (reanalyzing data from number of studies and finding lack of robustness to, for example, exclusion of influential outliers). Kang et al. do not mention or address these early challenges. Note that one of the authors of this piece (Kang) wrote an updated 2021 article directed towards judges that did not mention many of the topics we have discussed, such as publication bias, questionable research practices, and reasons effect sizes may not accumulate over time and people. *See* Kang, *supra* note 115, at 80. For a critique of the 2021 article focused on effect sizes, *see* Chin, *supra* note 116, at 2.

On one view of the scientific process, the epigraphs' divergence is business as usual. Science self-corrects and evolves.[196] Legal researchers, commentators, and others in the legal system do their best to rely on state-of-the-art scientific evidence. We take a somewhat different view. We believe that chalking up faulty reliance on implicit bias research in legal applications to those familiar processes is a lost opportunity to improve how science is used in law. Rather there have been questions for years about one of the key pillars of implicit bias, that subtle cues in the environment have meaningful effects on behavior that people are not aware of ("priming" effects). As noted, Daniel Kahneman recently disavowed this area of research altogether.[197] There have never been sustained attempts to test the generalizability of many implicit bias findings outside of university laboratory contexts. Put simply, the review of psychological scientific findings reflected in the first epigraph and other similar sources[198] reveals an uncritical and rhetorical account of implicit bias and its applicability to numerous legal contexts. What exactly went wrong with the science and how scientists communicated their findings, and what does it tell us about the path forward?

In what follows, we offer a brief critical review of implicit bias research and its applicability to discrimination in law. To be clear, we are not suggesting that people's intentions do not sometimes diverge from what they do. And, we emphasize that discrimination, lack of diversity, and prejudice are serious problems in the legal system and society more broadly. Rather, our concern stems from a desire to help legal actors avoid expending resources on legal reforms that have little chance of affecting human behavior because they are designed around unreliable scientific evidence.

Our focus in this Part is on the methodological issues reviewed in Part I, such as publication bias, questionable research practices, and limited generalizability. Our emphasis on research credibility supplements previous research suggesting that the implicit bias narrative may be harmful in removing blame from actors[199] and in not being theoretically coherent.[200] Moreover, we are not asserting that poor research methods were a sufficient condition for what we see as the rhetorical use of implicit bias as a means to address serious problems. Rather, the social and political context that

---

[196] *But see* Simine Vazire & Alex O. Holcombe, W*here Are the Self-Correcting Mechanisms in Science?*, 26 REV. GEN. PSYCH. 212, 1 (2022) ("It is often said that science is self-correcting, but the replication crisis suggests that self-correction mechanisms have fallen short.").

[197] In a 2022 lecture, Kahneman seemed to disavow behavioral priming. Kahneman, *supra* note 57 ("The crisis has been great for psychology. In terms of methodological progress, this has been the best decade in my lifetime. Standards have been tightened up, research is better, samples are larger. People pre-register their experimental plans and their plans for analysis. And behavioral priming research is effectively dead.").

[198] *See infra* Part III.

[199] Michael Selmi, *The Paradox of Implicit Bias and a Plea for a New Narrative*, 50 ARIZ. ST. L.J. 193, 200 (2018) ("labelling behavior as implicit that could just as easily be described as explicit, and, by doing so, much of contemporary discrimination is likely to evade legal liability.").

[200] For example, it is not clear that implicit bias evades conscious introspection. Mitchell, *supra* note 192, at 40 ("[A] number of lines of evidence cast doubt on the view that implicit measures actually tap into unconscious bias.").

motivated research on implicit bias is another underrecognized but crucial step towards understanding its rise in science, the popular discourse, and in legal circles. We briefly review that context before turning to the methodological problems with implicit bias research.

A. An introduction to implicit bias

Implicit bias research arose in the United States after the Civil Rights Movement.[201] It can be seen as a distinctively American phenomenon because it was driven by the country's history with race relations and social scientific attempts to study those processes and to improve them. Some psychology researchers were optimistic that policy and law reforms would reduce racism, and this opinion was seemingly vindicated by increasingly positive attitudes toward Black people in national opinion surveys.[202] Others challenged that view: "Our own position has been to question the assumption that verbal reports reflect actual sentiments, and we inferred from the literature that whites today are, in fact, more prejudiced that they are wont to admit."[203] This view was consistent with the trend in social psychology at the time, which tended to distrust self-report measures of attitudes (i.e., explicit measures) in favor of studying what people actually do (i.e., behavioral measures).[204]

Behavioral priming studies (recall the old-age priming study that gave rise to psychology's replication crisis), therefore, were central to the claim that people could act in ways that diverged from their verbalized attitudes. Thomas Srull and Robert Wyer's influential study embodies this trend. They asked participants to unscramble sentences that either related to aggression or not, and then, in a separate task, judge whether a specific person was acting in a hostile way.[205] They conjectured that exposure to hostility and aggression concepts would "prime" participants to perceive other people as aggressive. The authors found what appeared to be a large effect. Hostility-primed participants rated the person as more hostile by 3 points on average on a 0 (no at all hostile) to 10 (extremely hostile) scale.[206]

---

[201] In this section, we draw and quote from a blog post published by one of the authors written as part of the writing process for the current Article. Patrick S. Forscher, *A Brief Intellectual History of Implicit Bias*, PERSISTENT ASTONISHMENT (Jan. 27, 2023), https://persistentastonishment.blogspot.com/2023/01/a-brief-intellectual-history-of.html

[202] Faye Crosby, Stephanie Bromley & Leonard Saxe, *Recent Unobtrusive Studies of Black and White Discrimination and Prejudice: A Literature Review*, 87 PSYCH. BULL. 546, 547 (1980) ("…many of the recent survey data indicate that antiblack prejudice among whites is less prevalent today than in the past.").

[203] *Id*. at 557.

[204] Roy F. Baumeister, Kathleen D. Vohs & and David C. Funder, *Psychology as the Science of Self-Reports and Finger Movements: Whatever Happened to Actual Behavior?*, 2 PERSP. PSYCH. SCI. 396, 396 (2007) ("For decades now, psychology students have been taught from the first day of class that psychology is the science of behavior and that its ultimate goal is to describe and explain what people do.").

[205] Thomas K. Srull & Robert S. Wyer, *The Role of Category Accessibility in the Interpretation of Information About Persons: Some Determinants and Implications*, 37 J. PERSONALITY & SOC. PSYCH. 1660 (1979).

[206] We are simplifying the paradigm here. The authors averaged several 0-10 scales using words related to hostile (e.g., unfriendly). *See id*. at 1664.

A meta-analysis of several similar studies seemed to confirm the effect's robustness.[207] This general line of thinking–that automatic mental associations could fundamentally color our decisions in ways we were not aware of or paying attention to–would greatly influence the development of implicit measures in the study of racial bias going forward.

Using a version of the Srull and Wyer paradigm, with words stereotypically associated with Black people as the prime, Patricia Devine kicked off modern research on implicit bias.[208] Specifically, she reported that priming participants with those stereotypic words produced more hostile evaluations of ambiguous behaviors. [209] In other words, Devine suggested that some stereotypes were "automatic,"[210] being acquired from repeated pairings of a social group with negative information picked up from the social environment. These stereotypes were characterized as conceptually distinct from those that are "controlled" via cognitive effort.[211] This bifurcation of stereotypes and their underlying cognitive processes was likely appealing, partly because it was optimistic (people may want to do the right thing, they just mess up along the way), partly because it provided a road toward intervention (reduce the influence of stereotypic associations), and partly because it had resonance: it reflected longstanding conflicts in American history. Devine, however, offered no direct measure of the automatic associations; she merely attempted to document their influence.

This changed with the introduction of implicit measures, including the much-discussed Implicit Association Test (IAT). First Russell Fazio and colleagues,[212] and then Anthony Greenwald and colleagues,[213] introduced what they claimed to be direct measures of automatic associations, what later came to be known as implicit bias. Taking the IAT involves categorizing words or images, with experimenters measuring participants'

---

[207] Jamie DeCoster & Heather M. Claypool, *A Meta-Analysis of Priming Effects on Impression Formation Supporting a General Model of Informational Biases*, 8 PERSONALITY & SOC. PSYCH. REV. 2, 9 (2004).

[208] Patricia G. Devine, *Stereotypes and Prejudice: Their Automatic and Controlled Components*, 56 J. PERSONALITY & SOC. PSYCH. 5, 10 (1989) ("Subjects were told that the experimenter was interested in how people form impressions of others. They were asked to read a paragraph describing the events in the day of the person about whom they were to form an impression. This paragraph is the now familiar 'Donald' paragraph developed by Srull and Wyer . . . "); Kang et al., *supra* note 2, at 1136 n.34, rely on this study, even suggesting a dose-response relationship ("In a seminal paper, Patricia Devine demonstrated that being subliminally primed with stereotypically 'Black' words prompted participants to evaluate ambiguous behavior as more hostile […] Those who received a heavy dose of priming (80 percent stereotypical words) interpreted a person's actions as more hostile than those who received a milder dose (20 percent).").

[209] Sample sizes were small and reported p-values hovered around 0.05. *See* Devine, *supra* note 208, at 11-12.

[210] *Id*. at 6.

[211] *Id*. at 6.

[212] Russell H. Fazio, Joni R. Jackson, Bridget C. Dunton, & Carol J. Williams, *Variability in Automatic Activation as an Unobtrusive Measure of Racial Attitudes: A Bona Fide Pipeline?*, 69 J. PERSONALITY & SOC. PSYCH. 1013 (1995).

[213] Anthony G. Greenwald, Debbie G. McGhee & Jordan K. L. Schwartz, *Measuring Individual Differences in Implicit Cognition: The Implicit Association Test*, 76 J. PERSONALITY & SOC. PSYCH. 1464 (1998).

reaction times as they sort.[214] For example, the IAT may present "white/good" on the left side of the screen and "bad/black" on the right side. Then, in the middle of the screen, faces of Black or White people or positively or negatively valenced words appear for the participant to categorize. If the center image is a White face or a positively valenced word, they are (for example) asked to press "e," which is on the left side of the keyboard. If it is a Black face or a negatively valenced word, they are to press "i." After doing that several times, the study switches such that black is paired with good and white is paired with bad.

If participants are slower when asked to categorize faces and words in the black/good and white/bad trials (the stereotype-inconsistent trials), proponents of the IAT suggest that this indicates an implicit bias against Black people.[215] And, indeed, results from IATs and analogous tests suggest that many experiment participants show some level of implicit racial bias.[216]

Outside of the scientific sphere, IAT's inventors frequently pitched it (and the concept of implicit bias) to the public as a useful tool for diagnosing and understanding serious societal ills.[217] Legal commentators perhaps unsurprisingly then would both encounter such discussions and view implicit bias as a promising way to better understand and reduce discrimination in legal contexts. We suspect that psychologists initially found implicit bias attractive for many reasons that make it attractive to legal scholars—it suggests optimistically that some racism is not intentional and offers a target for reducing racism. Indeed, we will see in Part III that many contemporary legal commentators suggest reducing implicit bias to reduce racism and discrimination.

---

[214] What follows is a simplified picture of the IAT. For a fuller description, *see* Mitchell, *supra* note 192, at 32-33.

[215] *See* Kang, et al., *supra* note 2, at 1130-31.

[216] *Id*. at 1130-31. Although we will focus on the methodological issues raised *infra* Part I, note that the serious conceptual issues with implicit bias and the IAT have always existed and, if anything, have intensified. Many were reviewed in the second epigraph that started this Part. Gawronski, *supra* note 34, at 574; *see also* Hart Blanton & James Jaccard, *You Can't Assess the Forest if You Can't Assess the Trees: Psychometric Challenges to Measuring Implicit Bias in Crowds*, 28 INT'L J. ADVANCEMENT PSYCH. THEORY 249, 249 (2017) (arguing that the IAT predicts neither individual nor group level discrimination well). The relationships between implicit measures and criterion measures were also assumed to indicate that implicit bias causes behavior. Yet these relationships could indicate that attitudes (not just implicit, but all attitudes) cause behavior, or they could indicate that people who discriminate against Black people come to hold negative attitudes. Even the assumptions about discrimination that underlies the theory of implicit bias–that it's widespread and done by everyone–has been questioned. Mitchell R. Campbell & Markus Brauer, *Is Discrimination Widespread? Testing Assumptions About Bias on a University Campus*, 150 J. EXPERIMENTAL PSYCH. GEN. 756, 756 (2021). Finally, accumulating evidence suggests that the developmental story behind implicit bias—that it's an automatic attitude acquired over time without awareness—cannot be true. Olivier Corneille & Gaëtan Mertens, *Behavioral and Physiological Evidence Challenges the Automatic Acquisition of Evaluations*, 29 ASS'N PSYCH. SCI. 531, 573 (2020) (reviewing research finding that automatic associations are malleable).

[217] Sometimes the IAT's inventors' published research diverged from what they were communicating to the public. *See* Jesse Singal, *The Creators of the Implicit Association Test Should Get Their Story Straight,* INTELLIGENCER (Dec. 5, 2017), https://nymag.com/intelligencer/2017/12/iat-behavior-problem.html.

B. Is implicit bias a useful target in reducing discriminatory behaviors in law?

We now turn to limitations in the evidence base for implicit bias that call into question its role in driving discrimination and the utility of reforms aimed at reducing implicit bias. We highlight issues such as publication bias and the failure of key studies to replicate when tested using registered research designs employing larger sample sizes. These limits do not appear to be regularly mentioned by influential legal commentators who seek to bridge the gap between social science and legal contexts.[218] We focus on the purported relationship between implicit bias and behavior because behaviors are the outcome most of interest in law.

Our analysis starts by revisiting hostility priming, one of the key findings and paradigms in the implicit bias narrative, with particular focus on the failure to replicate a foundational study. Then, we turn to claims that implicit bias in the legal system can be reduced and that such reductions have meaningful behavioral effects. Such interventions include exposing participants to counterstereotypic examples of individuals from stigmatized groups[219] and training participants with plans for how to think and behave when they encounter someone from a stigmatized group.[220] As with priming findings, research on interventions is plagued by publication bias, is opaquely reported, and is conducted on undergraduate students in scenarios that are not obviously good proxies in applications to legal contexts. These limits are in many cases severe and should be acknowledged. Part III assesses whether legal scholars acknowledge such limits.

First, recall priming effects, which support a foundational assumption of implicit bias and its application to law. Priming studies purport to show that automatic (versus controlled) stereotypes affect behavior in a way people cannot easily introspect about and control. Legal commentators also rely on these studies to explain how implicit bias may impact legally relevant behaviors in ways that actors are unaware of (and thus should be reduced through training programs).[221]

---

[218] Despite claims like Kang et al. *supra* note 2, at 1136 fn.34, there has been some progress Recently, the Australian Law Reform Commission acknowledged the lack of support for implicit bias interventions. AUSTRALIAN L. REFORM COMM'N, WITHOUT FEAR OR FAVOUR: JUDICIAL IMPARTIALITY AND THE LAW ON BIAS, ALRC REPORT 138 131 (Dec. 2021) ("A number of strategies directed at the individual level, and employed on a widespread basis, have now been shown to be largely ineffective at changing behaviour.").

[219] *See* Nilanjana Dasgupta & Anthony G. Greenwald, *On the Malleability of Automatic Attitudes: Combating Automatic Prejudice with Images of Admired and Disliked Individuals*, 81 J. PERSONALITY & SOC. PSYCH. 800, 803 (2001).

[220] *See* Saaid A. Mendoza, Peter M. Gollwitzer & David M. Amodio, *Reducing the Expression of Implicit Stereotypes: Reflexive Control Through Implementation Intentions*, 36 PERSONALITY & SOC. PSYCH. BULL. 512, 513-514 (2010).

[221] For an example of reliance on priming studies, see, e.g., Kang et al., *supra* note 2, at 1129 ("How have mind scientists discovered such findings on matters so latent or implicit? They have done so by innovating new techniques that measure implicit attitudes and stereotypes that by definition cannot be reliably self-reported. Some of these measures involve subliminal priming and other treatments that are not consciously detected within an experimental setting. Other instruments use reaction time differences between two types of tasks-one that seems consistent with some bias,

Evidence from psychology's credibility revolution suggests that many behavioral priming studies, including the Srull and Wyer hostility study pivotal to implicit bias's history, likely report either false positive findings or exaggerated effect sizes. Most notably, a large replication (published in 2018 and including over 7,300 participants) using registered methods repeating the Srull and Wyer hostility study design found no effect of priming. [222] This failed replication and others contributed to Kahneman's conclusion in 2022 that: "The crisis has been great for psychology. In terms of methodological progress, this has been the best decade in my lifetime. Standards have been tightened up, research is better, samples are larger. People pre-register their experimental plans and their plans for analysis. **And behavioral priming research is effectively dead**."[223]

If psychologists cannot activate a concept like hostility outside someone's awareness and consistently observe a behavioral effect, then it becomes less plausible that measures like the IAT can meaningfully predict behavior. In other words, the IAT was built on the assumption that some people have particularly strong automatic associations between concepts like Black and hostility. However, if priming (e.g., activating concepts like race and hostility) does not affect judgments and behavior, it seems unlikely that changing people's automatic associations will be useful in reducing either those judgments and behavior, or tendencies that should be even harder to change, such as discrimination in legal judgments. Stated differently, if automatic associations do not predict behavior, then attempting to change someone's IAT score to change their behavior–if that is possible–seems futile.

Indeed, the scientific basis for the claim that IAT scores predict behavior has shown to be weaker than once thought. One of the key original studies relied on only 42 White university students,[224] and a re-analysis found that the results hinged on several methodological choices that seem difficult to justify, such as including one outlier that significantly affects the

---

the other inconsistent-as in the Implicit Association Test (IAT)."); For other examples, *see* Elizabeth Thornburg, *(Un)Conscious Judging*, 76 WASH. & LEE L. REV. 1567, 1626 (2019); Mikah K. Thompson, *Bias on Trial: Toward an Open Discussion of Racial Stereotypes in the Courtroom*, 2018 MICH. ST. L. REV. 1243, 1267-68 (2018).

[222] Randy J. McCarthy et al., *Registered Replication Report on Srull and Wyer (1979)*, 1 ADV. METH. & PRAC. PSYCH. SCI. 321, 324 (2018); *see also* Randy J. McCarthy et al., *A Multi-Site Collaborative Study of the Hostile Priming Effect*, 7 COLLABRA: PSYCH. 18738 (2021), https://pure.coventry.ac.uk/ws/portalfiles/portal/41835886/A_multi_site_collaborative_study.pdf (manuscript at 3) ("Despite our best efforts to produce favorable conditions for the effect to emerge, we did not detect a hostile priming effect."). For a review of behavioral priming studies, *see generally* Erik Mac Giolla, Simon Karlsson, David A. Neequaye & Magnus Bergquist, *Evaluating the Replicability of Social Priming Studies*, PSYARXIV PREPRINTS (July 8, 2022), https://psyarxiv.com/dwg9v/. A recent commentary on implicit bias written for a judicial audience does not acknowledge that behavioral priming effects do not replicate. *See* Kang, *supra* note 115, at 78-79. This is despite earlier research relying heavily on the priming narrative. *See* notes at *supra* note 221.

[223] Kahneman, *supra* note 57 [**emphasis added**].

[224] Allen R. McConnell & Jill M. Leibold, *Relations Among the Implicit Association Test, Discriminatory Behavior, and Explicit Measures of Racial Attitudes*, 37 J. EXPERIMENTAL SOC. PSYCH. 435, 436 (2001).

result.[225] Moreover, even if IAT scores are causally connected to behavior, little research suggests that IAT scores can be changed in a meaningful and enduring way. For instance, one influential study found that exposure to counter-stereotypic examples (with IAT changes assessed both immediately and 24 hours later),[226] saw its effect size reduced to 10% of the original estimate (from d = 0.82 in the original study to an average, weighted by sample size, of d = 0.08 in replications) when measured in a replication using a larger sample (about 4,000 participants) even when participants' IAT scores were measured just after the intervention aimed at reducing them.[227]

The growing body of evidence against claims that implicit bias measures are associated with (let alone have a causal relationship with) discriminatory behavior and that interventions can meaningfully change implicit bias measures compelled two research teams to conduct systematic reviews of evidence related to interventions designed to reduce bias and

---

[225] Hart Blanton, James Jaccard, Jonathan Klick, Barbara Mellers, Gregory Mitchell & Philip Tetlock, *Strong Claims and Weak Evidence: Reassessing the Predictive Validity of the IAT*, 94 J. APPLIED PSYCH. 567, 571-73 (2009); Mitchell, *supra* note 192, at 45-55, recently reviewed research finding that the correlation between IAT scores and behaviors estimated by meta-analyses were in the range of $r = 0.10$ to $r = 0.24$. A correlation coefficient of $r = 0.10$ implies that variation in implicit bias explains about 1% of the variation in discriminatory behavior ($r = 0.24$ corresponds to about 6%). Taking the square of a correlation coefficient (r) converts it to a coefficient of determination, R^2, which is a measure of the percentage of variation of the outcome variable (in our case, discriminatory behavior) that can be explained by variation in the explanatory variable (implicit bias). Daniel J. Ozer, *Correlation and the Coefficient of Determination*, 97 Psych. Bull. 307, 307 (1985) ("It is common practice in psychological statistics to use the square of the correlation as a coefficient of determination or a percentage measure of variance accounted for.") Thus, $0.10^2 = 0.01$ (or 1%) and $0.24^2 = 0.058$ (or roughly 6%). Beyond quantifying effect sizes, those suggesting implicit bias has a meaningful effect on legal outcomes should also justify why that effect size is practically meaningful with verifiable lines of reasoning, *see generally* Farid Anvari, Rogier Kievit, Daniël Lakens, Charlotte R. Pennington, Andrew K. Przybylski, Leo Tiokhin, Brenton M. Wiernik & Amy Orben, *Not All Effects Are Indispensable: Psychological Science Requires Verifiable Lines of Reasoning for Whether an Effect Matters*, PERSP. PSYCH. SCI (2022).

[226] Dasgupta & Greenwald, *supra* note 219 at 802-803.

[227] We are relying on Joy-Gaba and Nosek's recalculation of effect sizes. Jennifer A. Joy-Gaba & Brian A. Nosek, *The Surprisingly Limited Malleability of Implicit Racial Evaluations*, 41 SOC. PSYCH. 137, 137 (2010). ("Compared to a control group, participants viewing admired Black and disliked White individuals showed less implicit preference for White people compared to Black people (d = .82). The effect persisted in a follow-up test 24 h later (d = .71)."). *Id*. at 137. The replicators, across several conditions, found a much smaller effect ("Further, while DG reported a large effect of exposure on implicit racial (and age) preferences (d = .82), the effect sizes in our studies were considerably smaller. None exceeded d = .20, and a weighted average by sample size suggests an average effect size of d = .08, or d = .10 for just Experiments 2a, 2b, and 3 – the replicable paradigm."). To Kang and colleagues' credit, they proactively noted this decrease in effect size. Kang et al., *supra* note 2, at 1172 ("Recent research has found much smaller debiasing effects from vicarious exposure than originally estimated."). For a more recent study, *see* Calvin K. Lai & Jaclyn A. Lisnek, *The Impact of Implicit Bias-Oriented Diversity Training on Police Officers' Beliefs, Motivations, and Actions*, PSYARXIV PREPRINTS (Feb. 7, 2023), https://psyarxiv.com/dxfq6/ (manuscript at 11) ("…we tested a day-long implicit bias-oriented diversity training that sought to increase U.S. police officers' knowledge of biases, concerns about bias, and use of evidence-based strategies to mitigate bias (total N=3,764). Relative to baseline, the training was immediately effective at increasing knowledge about bias, concerns about bias, and intentions to address bias. However, the effects were fleeting […] These findings suggest that diversity trainings as they are currently practiced are unlikely to change police behavior.").

discrimination.[228] Elizabeth Paluck and colleagues' review was broader, examining interventions designed to change explicit as well as implicit measures of bias and related behaviors,[229] whereas Patrick Forscher and colleagues' review focused only on implicit bias.[230] The reviews agree that serious deficiencies plague the evidence base for all studied implementations.

Both reviews found considerable publication bias and other issues with research reporting and transparency. Paluck and colleagues, for instance, found a "powerful" relationship between estimates' precision and magnitude, suggesting that smaller, less precise estimates were driving aggregate estimates reported in meta-analyses. As noted above, smaller and null effects in larger studies are a sign of publication bias because it is possible to run many small studies and only publish those that "work," whereas it is less cost-effective to do so with studies employing large samples. Paluck et al. concluded that "if the current collection of studies had been conducted on a much larger scale, our analysis would have shown no reduction in prejudice."[231] This pattern was consistent across all types of prejudice reduction manipulations they studied: "in every theoretical domain we find unmistakable indications of publication bias: Large-N [i.e., large-sample] lab, online, or field studies that generate precise results tend to produce much weaker effects."[232] In their review of the implicit bias reduction literature, Forscher and colleagues also found evidence of publication bias.[233]

Regarding reporting and transparency within studies, Paluck and colleagues found their systematic review challenging because of poor reporting practices: "we encountered widespread lack of transparency insofar as few studies made their data and code publicly available. Many studies neglected to report key statistics such as standard errors or selectively reported estimated treatment effects only for subsets of the data, and the lack of public data made it impossible for us to calculate the relevant statistics ourselves."[234] As discussed, opacity in research practices allows questionable research practices to go undetected, which, in turn, contributes to false positives and inflated effect estimates.

---

[228] Forscher et al., *supra* note 67, at 522; Paluck et al., *supra* note 67, at 533.
[229] Paluck et al., *supra* note 67, at 536 ("Theoretically, we used a definition of prejudice that would encompass standard usage by researchers working in this domain: Prejudice is animus, or negative bias, toward social groups and their putative members. To be selected, the studies also needed to research an intervention that in some way sought to reduce prejudice as a psychological predisposition or its expression in behavior or behavioral intentions.") In an earlier review of 985 studies conducted more than a decade ago, Paluck & Green, *supra* note 170, at 339, focused mostly on experiment design issues and concluded that more work was required to make claims about the efficacy of interventions ("…the causal effects of many widespread prejudice-reduction interventions, such as workplace diversity training and media campaigns, remain unknown.").
[230] Forscher et al., *supra* note 67, at 525.
[231] Paluck et al., *supra* note 67, at 538.
[232] *Id*. at 539.
[233] Forscher et al., *supra* note 67, at 538.
[234] Paluck et al., *supra* note 67, at 525.

Both reviews also flag considerable issues with generalizability. Paluck and colleagues found that only about 10% of studies in their sample measured behavioral outcomes, with the vast majority instead surveying participants about their attitudes towards marginalized groups.[235] This limits application to legal contexts because behavioral outcomes are often most relevant in such contexts. Moreover, fewer than 10% of the studies sampled adults outside the university context, with about two thirds using university student samples.[236] Field studies were also rare.[237] Forscher and colleagues' review of implicit bias interventions raised similar concerns with only about 20% of samples drawn from non-university student populations, 10% measuring outcomes over more than just one time period, and 20% measuring behavioral outcomes.[238] While these study characteristics are consistent with the state of psychological science generally, they raise serious concerns when the findings are used to support implementations outside the lab aimed at reducing racism and discrimination. And, as we will see in Part III, legal academics and others regularly propose bias reduction interventions to address social problems.

Finally, both reviews found little evidence for the effectiveness of interventions aimed at reducing implicit bias and related behaviors. Paluck and colleagues found that just 17% of their sample estimated effects on implicit measures.[239] The associated average effect size was $d = 0.35$ (see Part I.A on interpreting effect sizes), but they suggested that this estimate is biased because the studies with small samples showed effects about twice the size of the average effect. The authors concluded that "a fair assessment of [their] data on implicit prejudice reduction is that the evidence is thin. Together with the lack of evidence for diversity training, these studies do not justify the enthusiasm with which implicit prejudice reduction trainings have been received in the world over the past decade."[240] They recommended that it "may be time to ramp up our investigation of attempts to reduce implicit prejudice while pausing the application of implicit prejudice interventions."[241]

Forscher and colleagues came to a similar conclusion, finding small to moderate relationships between implicit bias reduction strategies and reduced implicit bias. They also assessed whether those changes in implicit bias measures led to reductions in discriminatory behaviors and found no effect.[242] A 2021 article by the lead author of the first epigraph that began this Part, directed at judges and cataloguing ways they can reduce implicit bias, cites neither the Paluck- nor the Forscher-led systematic reviews.[243]

---

[235] *Id*. at 540.
[236] *Id*. at 539.
[237] *Id*. at 540.
[238] Forscher et al., *supra* note 67, at 532-33.
[239] Paluck et al., *supra* note 67, at 549.
[240] *Id*. at 549.
[241] *Id*.
[242] Forscher et al., *supra* note 67, at 536.
[243] Kang, *supra* note 115, at 81 ("What to do about implicit bias? Some evidence-based recommendations.").

To summarize, the evidence linking implicit bias to behavior is weak; thus, the overall body of research does not support the contention that discriminatory behavior can be reduced by reducing implicit bias. This conclusion is inconsistent with the descriptions of implicit bias and concomitant recommendations of legal commentators summarized in this Section. The common legal view of implicit bias training as a quick fix for social change was predictable—it reflects a strong desire to address a monumentally important social problem. In some ways, it also mirrors the overly simplistic ways that IAT's creators pitched it in the media, suggesting the need for better scientific communication at various stages in the research-to-policy pipeline. The following Section provides guidance for legal researchers and other users of scientific research to determine whether a particular body of research is ready to be relied upon in applied contexts, and, if not, how to proceed.

C. Ripeness of science

We recognize that legal researchers and others may find it useful to cite and draw upon individual studies and often justifiably choose not to wait for large-scale registered replication projects that assess the credibility of reported results or follow-up studies that provide empirical evidence to support generalizations assumed in policy recommendations. In this case, the best course of action is to carefully qualify recommendations, explain all steps in required generalizations, and acknowledge the limits of any relied-upon research.[244] In addition, those who wish to import findings that have yet to undergo extensive pre- and post-publication peer review, including replication, can employ available techniques for assessing study credibility. The following are examples of such techniques that focus on observable signals of possible credibility issues. We also suggest guidelines for developing and discussing generalizability assumptions.

First, studies reporting *p*-values at or near 0.05 are especially at risk of failing to replicate.[245] Many older priming studies report *p*-values hovering around 0.05 on key analyses.[246] Consider, for instance, a study that primed participants with photos of White versus Black men. It found that that manipulation decreased and increased (respectively) participants' recognition of the outline of a gun in a distorted image.[247] The study, however, included only about 40 participants total (20 per condition) and on key tests, the authors reported a *p*-value just below 0.05 and 0.05 exactly.

---

[244] *See infra* Part III for a detailed discussion.

[245] Adam Altmejd, Anna Dreber, Eskil Forsell, Juergen Huber, Taisuke Imai, Magnus Johannesson, Michael Kirchler, Gideon Nave & Colin Camerer, *Predicting the Replicability of Social Science Lab Experiments*, 14 PLOS ONE 1, 11 (2019) ("The statistical properties (p-value and effect size) of the original experiment are the most predictive.").

[246] Devine, *supra* note 208, at 11; Jennifer L. Eberhardt, Phillip Atiba Goff, Valerie J. Purdie & Paul G. Davies, *Seeing Black: Race, Crime, and Visual Processing*, 87 J. PERSONALITY & SOC. PSYCH. 876, 880 (2004); Jeffrey J. Rachlinski, Sheri Lynn Johnson, Andrew J. Wistrich & Chris Guthrie, *Does Unconscious Racial Bias Affect Trial Judges*, 84 NOTRE DAME L. REV. 1195, 1214 n.94, 1216 n.98 (2008).

[247] Eberhardt et al., *supra* note 246, at 880.

As noted in Part I, such results raise red flags because a wider range of *p*-values is expected in the absence of questionable research practices. This is not to say that *p*-values around 0.05 with small sample sizes are a sure sign that a study is unreliable, but further evidence is required before such findings can be considered reliable.[248] Those who rely on such studies should warn readers and cite to studies demonstrating that the probability of non-replication increases with small sample sizes and *p*-values at or just below 0.05.

Second, analyses of experiment participants' responses to stimuli that do not account for random variation in stimuli samples might lead to false positives (i.e., concluding that effects exist when they, in fact, do not).[249] In the implicit bias literature, researchers regularly expose participants to photographs of faces of Black and White individuals, but then generalize responses to these specific faces to a broader population of interest (usually the faces of all Black and White people in the sampled population). If the researcher does not account for random variation in the stimuli samples, we should be cautious about relying on reported statistical significance levels. Those who rely on such results should warn readers about the potential for unreliable conclusions.

Similar issues arise when findings are generalized beyond populations from which participant samples are drawn. As with most psychology research, priming studies rely heavily on recruiting participants from Western, college student populations. Given that even seemingly less culturally sensitive research often depends on the characteristics of participants (e.g., the effects of observing mean energy usage of one's neighbors on one's own energy usage),[250] caution is required. It might be that well-conducted studies using Western college student populations underestimate the magnitude of biases that exist in broader populations of interest.

Finally, priming studies and implicit bias studies raise serious questions around generalizations to contexts outside the lab. Exposing participants to flashing images or words in a lab diverges substantially from legally relevant conditions outside the lab, where more contextual information about an observed person is usually available, and many other factors can influence the observer's reactions and behavior. Kang and his colleagues sought to assuage generalizability concerns by pointing to findings from field experiments. For example, some studies have found that resumés from individuals in marginalized groups are rated less favorably than equivalent resumés from White people.[251] Kang et al. argue "The studies by Marianne Bertrand and Sendhil Mullainathan demonstrating

---

[248] Altmedj et al., *supra* note 245, at 11.
[249] Judd et al., *supra* note 152, at 62-3 (using priming studies as examples and suggesting statistical analysis methods for properly computing statistical significance when stimuli samples drawn from larger populations); *see also* Yarkoni, *supra* note 30, at 2-4.
[250] Bryan et al., *supra* note 147, at 981.
[251] Kang et al., *supra* note 2, at 1155 ("field experiments have provided further confirmation under real world conditions.").

discrimination in callbacks because of the names on comparable resumes have received substantial attention in the popular press as well as in law reviews."[252] Such studies, however, cannot bolster the generalizability of laboratory studies because many reasons besides priming and implicit bias possibly explain the findings of resumés studies, such as explicit bias.

Given growing concerns about the reliability of priming and IAT studies, we would expect that anyone recently recommending an intervention aimed at reducing implicit bias and relying on this body of work would include warnings about lack of reliability and discuss the empirical findings that call such studies into question. In the next Part, we report results from a review of 100 law journal articles that reference "implicit bias training" to determine the warning rate.

III. Implicit bias training references in law journal articles

Given the uncertainties described above, it seems advisable that anyone recommending an intervention to reduce implicit bias would acknowledge, at a minimum, the limitations of research indicating that an intervention might work.[253] Such circumspection is especially important in law where, as Kang and colleagues noted, findings migrate from "science journals into the law reviews".[254] Courts and policymakers then rely on unreliable empirical research and faulty recommendations based on that research.[255] At the same time, these legal actors cannot be expected to have the expertise to critically appraise psychology claims.

In light of the potential dangers of uncritically recommending psychological interventions, we sought to more systematically study references to "implicit bias training" in recent law journal articles. We chose trainings, as opposed to more foundational research on implicit bias, because of the practical consequences of recommending interventions without signaling limits in their research base. For instance, training is not costless, and engaging in ineffective interventions may give a false sense that they are reducing discrimination. Authors, therefore, should reasonably be expected to acknowledge the known limits of implicit bias training.[256]

---

[252] *Id*.

[253] Anecdotally, legal writers at highly respected journals do sometimes fail to acknowledge the limits of psychological science. *See* Jason M. Chin, Simine Vazire, Crystal N. Steltenpohl, Tobias Heycke, David T. Mellor, Justin T. Pickett, Alexander C. DeHaven, Alex O. Holcombe & Kathryn Zeiler, *Improving the Credibility of Empirical Legal Research: Practical Suggestions for Researchers, Journals and Law Schools*, 3 LAW TECH. & HUMS. 107, 109 (2021).

[254] Kang et al., *supra* note 2, at 1126.

[255] State v. Plain, 898 N.W.2d 801, 841 (Iowa 2017); State v. Rashad, 484 S.W.3d 849, 860 (Mo. Ct. App. 2016) (Van Amburg, C.J., concurring); Chin et al., *supra* note 3, at 1141-1144.

[256] Another issue, beyond the scope of this study, is that there is no common definition of implicit bias training. Moreover, the "implicit bias" label may represent branding to the extent that organizations simply say their training addresses "implicit bias" in order to fit concerns in the zeitgeist. In any case, journal articles raising the idea that implicit bias training can be trained away should acknowledge limits in the research.

A. Methods

1. Overview and design

We conducted an observational study of 100 law journal articles mentioning "implicit bias training" published from 2017 to 2022. This study used a reflexive thematic analysis approach to better understand whether and how the articles in our sample were acknowledging the scientific limitations reviewed above.[257] Specifically, we read these articles and looked for themes (see below) that related to our question, such as whether the articles recommended implicit bias training and whether they expressed any skepticism about its effectiveness. The study was prospectively registered (https://osf.io/945ur/registrations) using Haven and colleagues' qualitative preregistration template.[258]

2. Identifying and screening articles

We built our sample by searching HeinOnline on January 11, 2022, with the following specifications. We searched HeinOnline's "Law Journal Library" database.[259] The list of journals and periodicals that the database included on our search date is available in our supplementary materials (https://osf.io/fgnbs). Using the HeinOnline advanced search page,[260] we searched "implicit bias training," restricting the dates to 2017-2022. In the "Section Types to Search" field we checked only "Articles" (e.g., not "Notes" or "Comments"). The search returned 229 articles, which we sorted by "Relevance." As registered (https://osf.io/945ur/registrations), we then performed our thematic analysis on the first 100 articles on that list. The raw download is available online (https://osf.io/4ke8d).

During the coding process, we encountered seven articles that we could not access. These were articles from trade journals (e.g., New York State Bar Association Journal) that were behind paywalls and unavailable to us.[261] The list of those inaccessible articles is available in our supplementary materials (https://osf.io/vqr98). We excluded those articles and replaced them with the next seven in our initial download. The final dataset of 100 articles is openly available (https://osf.io/8ebuw).

3. Developing themes

To develop the coding scheme, one author read the first 20 articles in the sample to determine why the articles mentioned implicit bias training

---

[257] Virginia Braun & Victoria Clarke, *Using Thematic Analysis in Psychology*, 3 QUALITATIVE RSCH. PSYCH. 77, 77 (2006).
[258] Tamarinde L. Haven, Timothy M. Errington, Kristian Skrede Gleditsch, Leonie van Grootel, Alan M. Jacobs, Florian G. Kern, Rafael Piñeiro, Fernando Rosenblatt & Lidwine B. Mokkink, *Preregistering Qualitative Research: A Delphi Study*, 19 INT'L J. QUALITATIVE METHODS 1, 1 (2020).
[259] We could not find a description of the Law Journal Library that was not behind a paywall, so took a screen capture. https://osf.io/34nxr (July 9, 2022).
[260] It is best practice to provide the full Boolean search string used, but HeinOnline does not support this.
[261] We searched using databases accessible by University of Sydney affiliates.

and how they expressed doubts about it if doubts were mentioned. Then, that author discussed those articles with two of the other authors. Through that discussion, they developed five general themes. Those themes are described in Table 2, along with examples from articles in the sample. We did not mean these themes to exhaustively describe how articles treat implicit bias training. Rather, we sought to explore whether or not articles recommend implicit bias training and whether they provide any warnings to the readers about the evidence base behind implicit bias training or at least note that it is not a panacea.

The first two themes are endorsements of implicit bias training. The first is an explicit recommendation and the second is an implicit recommendation, such as describing implicit bias in primarily positive terms or as an example that worked well in situations similar to the subject matter of the article (see examples in Table 2). The third theme encompasses explicit and/or implicit skepticism about implicit bias training, such as by saying it may not reduce implicit bias or by describing research that suggests implicit bias training may not work. The fourth theme includes statements saying that while implicit bias training may be useful, it should be accompanied by other reforms or interventions. Articles embody the fifth theme if they mentioned implicit bias training but did not recommend it, critique it, or cite research that was critical of it (i.e., it did not fall into any of the other themes). Examples of all these themes are in Table 2.

Table 2. Examples of themes

Themes	Examples
(1) Explicit recommendation: The article recommended implicit bias training or included it in a list of recommendations	● "Although jurors are not employed by the government, they too should be required to complete implicit bias training as part of jury duty. Not only will implicit bias training help ensure a fair outcome of the case, it would help build a more just society. The easiest, and perhaps most efficient way, to train jurors on implicit bias is to do so during jury selection."[262] ● "Further, all decision-makers who determine moral character and fitness should have implicit bias training and be educated on substance use disorder, mental health, and recidivism data."[263]
(2) Implicit recommendation: The article suggested the reader "consider" implicit	● "A successful training will challenge assumptions and 'get people to understand how other people perceive their actions.' Conducting implicit-bias training will demonstrate that the company is committed to inclusivity and will assist

[262] Olga M. Torres, *The Not-So-Hidden Bias Lurking Within the Criminal Justice System: Issues of Race and Gender in the Law*, 3 Soc. Just. & Equity L.J. 132, 160 (2019).
[263] Tarra Simmons, *Transcending the Stigma of a Criminal Record: A Proposal to Reform State Bar Character and Fitness Evaluations*, 128 Yale L.J.F. 759, 770 (2018).

bias training or described it in primarily positive terms	decision-makers on making pay equity decisions that are fair and objective."[264] ● "Most importantly, impartiality training would allow the courts to advertise to the public that judges not only understand the importance of divorcing personal beliefs from court decisions but also receive training how to accomplish that goal; for example, State v. Plain (2017: 841) explicitly states that all Iowa judges are required to undergo implicit bias training and testing."[265]
(3) Skepticism: The article expressed skepticism about implicit bias training, described research that demonstrated reasons for skepticism, or express skepticism about research supporting implicit bias training. Here, we mean skepticism about its effectiveness in reducing implicit bias and/or its effects. We included statements raising doubt about whether training is effective in the long run despite working for shorter periods of time	● "Programs designed to alleviate implicit biases, for example, require that political actors and the target audience accept the existence of such biases and the need to address them. Without such a recognition, implicit bias training may produce a backlash and be counterproductive."[266] ● "Often, prosecutors' offices start with the now-obligatory implicit bias training. While there is a place for such training in moving forward equity and racial justice, too many times the work ends there. A one-time course does little to undo years of learned behavior as well as unpack some of the drivers of inequality."[267]
(4) Incomplete: the article noted that implicit bias training may have some use, but that it is not a complete solution.	● "The Stop Militarizing Act attempts to eliminate tools of excessive destruction; implicit bias trainings try to cure the false tendency to view Black people as dangerous [...] But while police reform is necessary, it is not enough."[268] ● "Acknowledging one's own biases is a necessary first step. Court programs and service providers should require mediators to take the IAT and engage in other bias reduction

[264] Christine Lyman, Lonnie Giamela & LaLonnie Gray, *Mind the Gap: Practical Solutions to Minimize Pay Equity Claims*, 49 COLO. LAW. 30, 39 (2020).
[265] Raymond J. McKoski, *When the Law and a Judge's Personal Opinions Collide*, *in* NCSC TRENDS IN STATE COURTS. 76, 80 (2020).
[266] Derrick Darby & Richard E. Levy, *Postracial Remedies*, 50 U. MICH. J.L. REFORM 387, 463 (2017).
[267] Melba V. Pearson, *Data as a Tool for Racial Justice*, 36 CRIM. JUST. 4, 4 (2021).
[268] Noa Ben-Asher, *Trauma-Centered Social Justice*, 95 TUL. L. REV. 95, 121 (2020).

	efforts to receive case referrals."[269]
(5) Mentioned: Implicit bias training is only mentioned or is found in the title of a reference	• "The most reported police practices were having a crisis intervention team, using cite and release in lieu of arrest, community engagement activities, and implicit bias training…"[270] • "166. See State v. Plain, 898 N.W.2d 801, 841 (Iowa 2017) (explaining that all Iowa judges are required to "undergo implicit-bias training and testing"); see also State v. Rashad, 484 S.W.3d 849, 860 (Mo. Ct. App. 2016) (Van Amburg, C.J., concurring) (indicating that the judicial education curriculum in Missouri now includes implicit bias training)…"[271] (footnote)

Table caption. The five themes extracted from articles in the sample and example quotes for those themes.

4. Data extraction procedure

Two authors independently extracted data from (i.e., "coded") the 100 articles. On a Google Sheet, they indicated "yes" or "no" to whether one of the five themes was present. Except for the following, the themes were not coded as being exclusive of other themes. First, articles that only mention implicit bias training were not coded as any other theme. Second, articles with explicit recommendations of implicit bias training were not coded as containing implicit recommendations. In other words, articles with explicit recommendations may have also included implicit recommendations.

After that process, one author identified disagreements and sent those articles along with a data extraction form (https://osf.io/k8dp3) to a third author to resolve the disagreements. That third coder was not told what themes caused the disagreement. If the third coder agreed with one of the other two coders on all themes, that agreed-upon set of coding was accepted. If not, two authors discussed the article until they agreed on all themes. This system of resolving disagreements was not registered, but all disagreements are reported in our data (https://osf.io/8ebuw). Overall, there were disagreements between coder 1 and coder 2 in 20 of the 100 articles. Of those 20 disagreements, 15 required discussion.

B. Results

The quantitative analysis below was performed in R Studio (2022.12.0). The code that cleans the data and performs the analyses is

[269] Izumi, *supra* note 39, at 690.
[270] Pamela K. Lattimore, Stephen Tueller, Alison Levin-Rector & Amanda Witwer, *The Prevalence of Local Criminal Justice Practices*, 84 FED. PROB. 28, 30 (2020).
[271] Christian B. Sundquist, *Uncovering Juror Racial Bias*, 96 DENV. L. REV. 309, 330 (2019).

openly available (https://osf.io/z4qf3). 95% confidence intervals were calculated using the Sison-Glaz method for multinomial proportions.[272]

The 100 articles in our dataset are spread across the years included in the sample, however, there are more articles in 2020 and 2021 than the earlier years (Table 3).[273]

Table 3. Sampled articles by year

Year	Number of articles
2017	12
2018	20
2019	13
2020	25
2021	30

Table caption. Number of articles (N = 100) by publication year. Note that some journals reported year ranges (e.g., 2020-2021). In cases like those, the article is classified in the later year of the range.

Table 4 presents the percentage of articles containing each of the five themes. As can be seen, explicit recommendations of implicit bias training represent the most common theme present in the sample, appearing in 45% of articles (95% CI = [36%, 56%]). Following that, mere mentions (26%, 95% CI = [18%, 35%]) and noting that implicit bias training is an incomplete solution (26%, 95% CI = [18%, 35%]) were next most common. Fewer implicitly recommended implicit bias training (13%, 95% CI = [7%, 19%]) or were skeptical of it (19%, 95% CI = [12%, 27%]).

[272] Cristina P. Sison & Joseph Glaz, *Simultaneous Confidence Intervals and Sample Size Determination for Multinomial Proportions*, 90 J. AM. STAT. ASS'N 366, 366-368 (1995).
[273] Note that some journals reported year ranges (e.g., 2020-2021). In cases like those, we report the later year in the range.

Table 4. Themes present in the sample

Theme present	Percent (95% CI)
Explicit recommendation	45% [36% to 56%]
Implicit recommendation	13% [7% to 19%]
Skeptical	19% [12% to 27%]
Incomplete solution	26% [18% to 35%]
Mentioned	26% [18% to 35%]

Table caption. The percent of articles falling within each theme, along with the 95% confidence interval for that percent. Except for the following, the themes were not coded as being exclusive of other themes. First, articles that only mentioned implicit bias training were not coded as any other theme. Second, articles with explicit recommendations of implicit bias training were not coded as containing implicit recommendations. In other words, articles with explicit recommendations may have also included implicit recommendations.

From these general results, we drill down into the 58 articles that explicitly or implicitly recommended implicit bias training to determine whether these articles included language mitigating those recommendations (Figure 1). We found that a minority of them expressed skepticism (14%, 95% CI = [7%, 23%]) and/or indicated that implicit bias training is an incomplete solution (33%, 95% CI = [22%, 46%]).

Figure 1. Percent of articles expressing skepticism or that implicit bias training is not a panacea within articles recommending implicit bias training

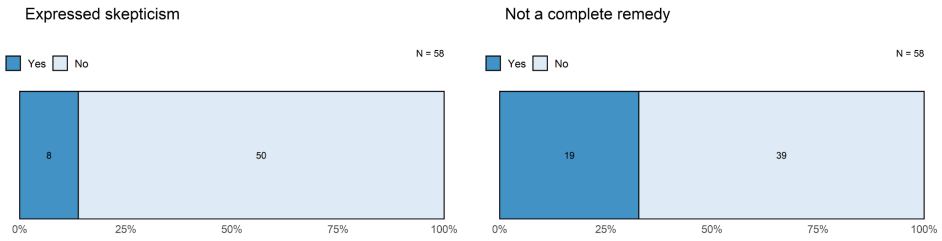

Figure caption. Among the articles recommending implicit bias training (N = 56, either explicitly or implicitly), the percent of those articles expressing skepticism of implicit bias training (left) or expressing that it is not a complete remedy (right).

C. Discussion

Our results are alarming in that only 19 of the 100 articles in our sample provide explicit or implicit skepticism about whether implicit bias

training works. This pattern did not improve when we limited our analysis to the 58 articles that recommended implicit bias training. These findings are troubling because, as we have discussed, the evidence for the effectiveness of implicit bias training program is limited at best, and, as we demonstrated, this was known well before these 100 articles were drafted.[274] Moreover, these programs are not costless and may provide a false sense of security that bias, implicit or otherwise, has been mitigated. It may be better to utilize interventions with a stronger evidence base.[275]

Uncritical recommendations of implicit bias may also adversely affect other scholarly articles in the legal-psychology literature. A failure to cite evidence on both sides of an issue, such as both against and for the effectiveness of implicit bias training, is sometimes referred to as "citation bias."[276] In various fields, citation bias combined with publication bias and other reporting biases yields very inaccurate literatures.[277] These biases can then snowball as subsequent papers add citation bias when citing older papers that themselves suffer from citation bias. The same likely is occurring in psychology and law for findings about implicit bias and implicit bias training, with chains of papers recommending implicit bias training with little or no reference to the studies finding it has null, small, or negative effects.

IV. How to address the limits of psychology in law

As we have seen, experimental psychology, like any field of research, is subject to a host of uncertainties and limitations. In exploring these over a century ago, Wigmore was inspired by what can now be clearly seen as overclaiming–a psychologist saying he could determine if a confession was true based on how quickly the accused associated words related to the crime versus irrelevant words. Wigmore asked how precise these methods were and if they had reached general acceptance in psychology. Since then–and especially in the past decade–the conversation in psychology has shifted, with metaresearch exploring more subtle ways in which psychology research can fail to be publicly useful.

Parts II and III of this paper revealed what can go wrong when this emerging body of metaresearch is ignored. In particular, implicit bias research is plagued by publication bias, many studies are not clearly applicable to realistic contexts, and the practical significance of effect sizes are not discussed appropriately in the literature. Articles for legal audiences rarely acknowledge these limits, making it difficult for such readers to

---

[274] *See supra* Part I. The credibility crisis in experimental psychology dates back to at least 2012 when Daniel Kahneman made his famous call for improvement. *See* Ed Yong, *Nobel Laureate Challenges Psychologists to Clean Up Their Act*, NATURE (Oct. 3, 2012), https://www.nature.com/articles/nature.2012.11535.

[275] *See* Forscher & Devine, *supra* note 33, at 303-304.

[276] Y. A. de Vries., A. M. Roest, Peter de Jonge, Pim Cuijpers, M. R. Munafò & J. A. Bastiaansen, *The Cumulative Effect of Reporting and Citation Biases on the Apparent Efficacy of Treatments: The Case of Depression*, 48(15) PSYCH. MEDICINE 2453, 2453 (2018) (noting that studies with positive results are cited more than those reporting null findings).

[277] *Id*.

apprehend and account for these substantial–perhaps fatal–limits. So, what can be done? We break our suggestions down based on the actor: psychology researchers, those in the legal community involved in communicating psychology research, and the users of psychology research (e.g., courts and policymakers).

A. Psychology Researchers

Psychology researchers bear a great deal of the responsibility in ensuring that research reaches users with the appropriate cautions. Despite this, discussion sections in psychological research often do not flag key limitations,[278] such as cautions about statistical conclusions and limited generalizability to contexts of interest to legal decision makers. As Phoebe Ellsworth put it in an address to the Association for Psychological Science in 2016, "We can assume that our colleagues will be skeptical about our claims in the discussion section and they'll evaluate them in terms of what they read in the methods sections and the results that we actually got. People that are not scientists–legislators, judges, reporters, the public–are more likely to read only the introduction and discussion."[279]

While much can be said about what legal-psychology researchers can do to avoid these pitfalls, we will focus on three. We will start with reporting research more cautiously and then move on to producing research that is *more useful* for legal applications and thus *less likely to be misused*.

First, we believe Ellsworth's call for increased circumspection among legal psychologists should be buttressed by more tangible guidelines about reporting limitations. As discussed, some guidance for experimental psychologists already exists, such as including "Constraints on Generality" statements in all articles.[280] In these statements, researchers explain why the study may or may not generalize to other contexts and populations. This is a salutary reform that appears to be catching on.[281] But, what seems to be missing is guidance specific to law and psychology. For instance, law and psychology is an applied field and so in many cases, it will be inappropriate to simply say that an effect was detected without explaining to readers whether and why that effect is practically relevant.[282] Legal commentators also do not always acknowledge that effect sizes reported in the implicit bias literature may be inflated due to publication bias and questionable research practices.[283] Perhaps if the relied-upon literature was more forthcoming about these factors, the resulting scientific communication

---

[278] *Infra* Part III.
[279] Ellsworth, *supra* note 28; The same applies when psychologists are preparing reports as expert witnesses. *See* Chin & Neal, *supra* note 45.
[280] Simons et al., *supra* note 43, at 1123.
[281] The article recommending constraints on generality has been cited 680 times. GOOGLE SCHOLAR, https://scholar.google.com/scholar?cites=10597154077887001511&as_sdt=2005&sciodt=0,5&hl=en (using "Constraints on Generality (COG): A Proposed Addition to All Empirical Papers" as search term).
[282] *See* Otgaar et al., *supra* note 121, at 4; *see also* Chin, *supra* note 116, at 2.
[283] *See* Chin, *supra* note 116, at 2.

would be more circumspect. Accordingly, we suggest that a sensible next step is for legal psychologists to develop guidelines or a standard format for limitations sections, perhaps with the limits reviewed in Part I as a starting point.

Beyond reporting, legal psychologists should think critically about how to plan and produce work that is more readily applied to pressing issues in law.[284] We acknowledge that resource limitations will always play a role. For instance, researchers often make tradeoffs, balancing sample size and the cost of recruiting participants. It is also often challenging to determine appropriate tradeoffs between similarity of study conditions to legally relevant situations and the need to maintain control over study conditions. Still, in some areas, more careful research planning and education of psychologists could assist and, in fact, ensure best use of existing resources. For instance, researchers could improve resource use by planning studies such that they are sufficiently powered to detect practically relevant or meaningful effects.[285]

Specifically, collaborating with research users may help researchers plan studies so that they are more applicable to reform efforts. For example, a researcher might survey stakeholders about minimum effect sizes that would be relevant for their decisions.[286] Researchers could then plan their studies to be able to detect effects of that size. Similarly, closer collaboration between end-users and researchers may be beneficial. We remarked above on the example of an Australian government-funded study into whether joining trials of multiple complainants leads to bias against the accused in those cases.[287] Commentators have pointed out, however, that the researchers did not plan the study carefully enough and so the null result may have been due to a too-small sample as opposed to the actual underlying effect being negligible.[288] In other words, the authors included enough mock jurors to find an effect size expected based on past research without consulting with legal stakeholders to determine if that effect size would be sufficient to warrant a change in the law.[289] In our view, this was a lost opportunity for stakeholders and researchers to come together to design a study that would usefully address a serious issue and potential change in the law.

---

[284] *Id*. at 3.
[285] *See* Otgaar et al., *supra* note 121, at 4.
[286] For instance, how many inconsistencies in a witness's testimony typically lead to a challenge to the admissibility of that evidence? Instead of surveys, these expert opinions may be extracted through structured expert judgment elicitation protocols. *See* Victoria Hemming, Terry V. Walshe, Anca M. Hanea, Fiona Fidler & Mark A. Burgman, *Eliciting Improved Quantitative Judgements Using the IDEA Protocol: A Case Study in Natural Resource Management*, 13 PLOS ONE 1, 1-2 (2018).
[287] For example, does joining trials cause mock jurors to conflate or double count evidence from the complainants? *See* Peter M. Robinson, *Joint Trials and Prejudice: A Review and Critique of the Report to the Royal Commission Into Institutional Child Sex Abuse*, 43 Monash U. L. Rev. 724, 728 (2018).
[288] Chin et al., *supra* note 3, at 1142-43; Robinson, *supra* note 287, at 736 ("which is true of jury verdicts because of the small sample size …")
[289] For a more detailed and technical explanation of this point *see* Chin, *supra* note 116, at 3.

Finally, legal-psychology researchers should think more carefully about the research to application pipeline.[290] Consider, by analogy, biomedicine, which is also an applied field. In that area, earlier stages of research (preclinical) start out by providing a proof of concept, often through testing an intervention with animal models. If that is successful, then the intervention is tested in a more tightly controlled study with humans. Indeed, clinical trials are preregistered by law in some jurisdictions[291] and there are attempts to quantify adverse drug effects.[292] While this process is far from perfect and has attracted substantial critical commentary,[293] the contrast with psychology research is stark. Notably, there is no formal research to action pipeline in legal psychology and, based on our review, much of this research is stuck in the proof-of-concept stage– the manipulation may work, but we do not know to what extent and under what conditions.

B. Law school administrators, law scholars, and law journal editors

To aid courts and policymakers in appropriately weighting evidence used to support normative and descriptive claims in law journal articles, those involved in the funding, production, and dissemination of legal scholarship can take a number of steps to provide information necessary for assessment. In addition, this group can work together to ensure that claims relying on faulty evidence do not make it into legal scholarship.

First, *funders* of legal scholarship have a role to play. The vast majority of legal scholarship is funded by law schools in the form of support for hiring law student research assistants. In addition, some law schools have added an empiricist to their staff to assist legal scholars in producing empirical legal research. Law school administrators have a few options for leveraging their funding position to improve the quality of law scholarship that relies on empirical findings. For one, they can hire staff-empiricists and require them to keep up with developments in metaresearch. They also can require relevant legal scholars to consult with the staff-empiricist as a funding condition. Alternatively, at a minimum, faculty should be encouraged to collaborate with the staff-empiricist.

An additional institutional approach might involve updating tenure standards. Some law school faculties require tenure-track faculty members

---

[290] For a proposal in psychology generally, *see* Hans IJzerman, Neil A. Lewis Jr., Andrew K. Przybylski, Netta Weinstein, Lisa DeBruine, Stuart J. Ritchie, Simine Vazire, Patrick S. Forscher, Richard D. Morey, James D. Ivory & Farid Anvari, *Use Caution When Applying Behavioural Science to Policy*, 4 NATURE HUM. BEHAV. 1092 (2020).

[291] Kay Dickersin & Drummond Rennie, *The Evolution of Trial Registries and Their Use to Assess the Clinical Trial Enterprise*, 307 J. AM. MED. ASS'N 1861, 1862 (2012).

[292] Bruno H Ch Stricker & Bruce M Psaty, *Detection, Verification, and Quantification of Adverse Drug Reactions*, 329 BMJ 44, 46-47 (2004).

[293] *See* Stylianos Serghiou, Cathrine Axfors & John P. A. Ioannidis, *Lessons Learnt From Registration of Biomedical Research*, 7 NATURE HUM. BEHAV. 9, 9 (2023) (claiming preregistration has been valuable in biomedicine but inconsistently applied). For the limits of measuring side effects, *see* Stricker & Psaty, *supra* note 292, at 44 ("…the current approach to this is scattered and disappointing").

to produce solely authored articles to earn tenure.[294] This makes it difficult for junior scholars who wish to rely on empirical evidence to collaborate with experts familiar not only with the relevant empirical methodologies but, ideally, also with the state of credibility of the field in which the evidence was produced. Law schools should not only allow juniors to engage in collaborative scholarship, but they also should encourage it, and possibly require it, whenever developing law scholars import empirical findings. In fact, teaming up with methodologists who are apprised of the state of credibility in their fields can help legal scholars, pre-tenure and tenured, avoid potential inferential pitfalls. To the extent that tenure standards allow limited exceptions to this requirement, they should count collaboration with empirical methodologists as a valid exception.

Second, *legal scholars*, themselves, can take steps.[295] In addition to teaming up with informed experts, legal scholars can make use of existing tools to conduct basic credibility checks on the studies they plan to rely on. Beyond merely understanding the full scope of the original empirical literature on a particular topic and avoiding cherry picking of individual studies or small groups of studies that support one's preferred findings,[296] importers of empirical studies must understand the state of credibility of the imported studies' fields.[297] At a minimum, authors should either clearly describe, in the text of both the introduction and conclusion sections, the state of the credibility of the relevant research fields based on the metaresearch that's been done in those fields or clearly state that readers should be aware that the fields might be facing general credibility issues and that all relied-upon findings are of questionable reliability.

Finally, *law journal editors*, including students who edit law journals housed in law schools, have a role to play as gatekeepers.[298] Editors

---

[294] For example, Boston University School of Law's tenure standards require production of at least two solely authored articles. An exception will be made for just one of the articles if it is an interdisciplinary work. Boston University Law School Faculty, Guidelines for the Appointment, Promotion and Tenure of Tenure Track Faculty 3 (Jan 10, 2023) (on file with author).

[295] Note that these prescriptions are related to the importation of empirical findings into legal scholarship. For an example of prescriptions for empirical legal scholars, *see* Bavli, *supra* note 65, at 504 (proposing "DASS" method–Design, then Analyze, then Scrutinize and Substantiate Adherence to method–to help empirical legal scholars avoid questionable research practice of p-hacking). For a broader set of guidelines, *see* Jason M. Chin, Alexander C. DeHaven, Tobias Heycke, Alex O. Holcombe, David T. Mellor, Justin T. Pickett, Crystal N. Steltenpohl, Simine Vazire & Kathryn Zeiler, *Improving the Credibility of Empirical Legal Research: Practical Suggestions for Researchers, Journals and Law Schools*, 3 LAW, TECH. & HUMS. 107, 116-120 (2021).

[296] *See* Kathryn Zeiler, *Cautions on the Use of Economics Experiments in Law*, 166 J. INSTITUTIONAL & THEORETICAL ECON. 178, 188-190 (2010). Some have suggested how search engines like Google Scholar might be improved to help researchers gain a more nuanced perspectives of relevant literatures. *See* Paul T. von Hippel & Stuart Buck, *Improve Academic Search Engines to Reduce Scholars' Bias*, NATURE HUM. BEHAV. (2023).

[297] For an example of an available resource, see, e.g., Garret Christensen, Zenan Wang, Elizabeth Levy Paluck, Nicholas Swanson, David Birke, Edward Miguel & Rebecca Littman, *Open Science Practices Are on the Rise: The State of Social Science (3S) Survey*, CEGA WORKING PAPER SERIES 12 (2019) (reporting estimates of use of open science practices across various fields over time); *see also* Vazire & Holcombe, *supra* note 196, at 4.

[298] Editors also act as gatekeepers when it comes to publishing only replicable empirical legal studies. For advice for journal editors related to this gatekeeping role, *see* Jason M. Chin & Kathryn

are well positioned to avoid publication of legal scholarship that relies on unreliable empirical studies in the absence of explicit warnings to readers. First, and perhaps ideally, the peer review process should always include an expert in metaresearch in the relevant fields. If that's not possible, editors should educate themselves about the on-going credibility crisis.[299] It's important for editors to understand the level of credibility of the relevant fields at the time of publication of the relied-upon studies. At a minimum, editors should employ a checklist to assess whether submitted articles have appropriately cautioned about the reliability of any inferences drawn from empirical studies and to determine whether the author has adequately considered indicia of *credibility* of each study cited including whether the study's author has ensured public access to the data, analysis scripts, and other research materials; posted a dated preregistration of the analysis plan prior to collecting data; explicitly justified sample sizes; and, explicitly reported any conflicts of interest and funding sources.[300] In addition, the article author should report, especially for experiments, a summary of the full set of replication attempts, if any, and whether the study's findings have been successfully replicated. This is especially important if an expert peer reviewer familiar with the metaresearch in the relevant fields has not reviewed the article. Requiring that all articles relying on empirical studies include a table in an appendix that provides information about indicators of credibility would go a long way to inform readers about the potential credibility issues.

C. Courts and policymakers

Finally, we address users of psychology research, such as courts and policymakers. Research users face a difficult task. Psychology research is not always conducted and reported in ways that clarify its limits. Research users are often not subject matter experts, they face their own resource constraints, and they often must make important decisions based on incomplete and uncertain information and research. We attempt to account for these challenges in our recommendations, which fall into two themes, how to treat psychology research with caution and the importance of consulting with methodology specialists (including metaresearchers) when appropriate.

Zeiler, *Replicability in Empirical Legal Research*, 17 ANN. REV. L. & SOC. SCI. 239, 254-255 (2021).

[299] Ample resources are available. *See e.g.*, JON GRAHE, A JOURNEY INTO OPEN SCIENCE AND RESEARCH TRANSPARENCY IN PSYCHOLOGY: A JOURNEY THROUGH NATIONAL PARKS (2021); MATTHEW C. MAKEL & JONATHAN A. PLUCKER, Toward A MORE PERFECT PSYCHOLOGY: IMPROVING TRUST, ACCURACY, AND TRANSPARENCY IN RESEARCH (2017); GARRET CHRISTENSEN, JEREMY FREESE & EDWARD MIGUEL, TRANSPARENT AND REPRODUCIBLE SOCIAL SCIENCE RESEARCH (2019).

[300] Balazs Aczel and colleagues provide a comprehensive list of transparency indicators. *See* Balazs Aczel et al, *A Consensus-Based Transparency Checklist*, 4 NATURE HUM. BEHAV. 4, 5 (2019).

First, research users should presume that many older psychology studies contain exaggerated results.[301] Part I *supra* surveys the basis for this caution, including the many large replication projects finding exaggerated claims and systematic studies comparing those large sample-sized, registered replications with smaller sample size studies from the years before. If courts, policymakers, and others must consider these older studies, they can use *indicia* of credibility, including the sample size and whether the study's main statistical conclusion is at or just below a *p*-value of 0.05. In addition, the Framework for Open and Reproducible Research Training (FORRT)[302] and ReplicationWiki[303] maintain living databases of social science studies that have failed to replicate and details of those failures.

Asking non-experts to critically review research is far from ideal. Accordingly, we also recommend that research users engage research specialists when necessary. It is already common for courts and policymakers to invite submissions from subject-matter experts.[304] Less commonly engaged, however, are experts that study research methods themselves. In our experience, this can prove helpful. For instance, two of the present authors consulted with the Australian Law Reform Commission (ALRC), the Australian federal body in charge of recommending legal reforms for the Parliament's consideration. The ALRC had published a preliminary report about implicit bias in law that included several studies of dubious credibility,[305] and indicated they were open to recommending implicit bias training for the Australian judiciary.[306] Based on submissions, including those from metaresearchers, the ALRC's final report acknowledged the limits of the implicit bias literature and ultimately did not recommend implicit bias training:

> [Submissions] emphasised that a significant amount of research in psychology and science more generally is currently subject to a 'replication crisis', and a process of

---

[301] Courts and policymakers should be especially cautious when legal scholars support normative claims by referring only to dated studies. Most fields of research, especially fields studying contentious topics, contain tens or hundreds of studies aimed at verifying and refining original findings. Once a sufficient number of studies have been published, researchers usually conduct meta-analyses to combine data to produce a single estimate or to assess the methodology of published studies and determine which are most reliable (e.g., which have sufficient power).
[302] Replications & Reversals, FORRT, https://forrt.org/reversals/ (last accessed Feb. 7, 2023).
[303] *ReplicationWiki*, REPLICATIONWIKI (last updated Aug. 29, 2022), https://www.semantic-mediawiki.org/wiki/ReplicationWiki (last accessed Feb. 9, 2023)
[304] For instance, in developing procedures for admitting eyewitness identification evidence, the New Jersey Supreme Court appointed a special master to hear from psychologists knowledgeable about eyewitness memory. For a review, *see* Amy D. Trenary, *State v. Henderson: A Model for Admitting Eyewitness Identification Testimony*, 84 U. COLO. L. REV. 1257, 1263 (2013). For expertise in matters of policy, *see* Chin et al., *supra* note 3, at 1148.
[305] AUSTRALIAN L. REFORM COMM'N, CONSULTATION PAPER: JUDICIAL IMPARTIALITY 6-9, 6-12 (Apr. 2021). For a review of the questionable studies referred to in that paper, *see* Jason. N. Chin, Letter to Australian Law Reform Commission (June 21, 2021), https://www.alrc.gov.au/wp-content/uploads/2021/07/14-.-Dr-Jason-Chin-Public.pdf
[306] AUSTRALIAN L. REFORM COMM'N, *supra* note 305, at 32.

> retesting widely cited studies is currently underway. Both stakeholders suggested that older research in particular should be treated with caution, although others noted that a number of tested heuristics and cognitive biases have been found to be robust. The analysis below recognises the potential limitations of some of the older research in this area. It attempts to capture the most relevant findings that have been validated as robust, or points to potential limitations of studies.[307]

We consider this to be a sensible approach that balances the importance of psychology research in informing policy with the reality that many psychology findings are not yet up to that task.

V. Conclusion: Towards a metaresearch agenda for law and psychology

The field of psychology has much to offer those interested in reforming law. Recent metaresearch findings, however, give legal scholars, judges, and policy-setting bodies ample reason to be cautious when applying findings from psychology experiments. When experiments that haven't yet been severely tested are used to bolster claims, it is critical that *indicia* of reliability be clearly communicated and appropriately used to weigh the strength of the evidence in support of or against reform proposals. In cases where they have been severely tested and have been found seriously lacking, alternative interventions must be considered, especially for vitally important issues like racism. Our analysis here should not be taken to imply that we should abandon implicit bias training wholesale, and it certainty should not be taken to imply that no effective interventions are possible. While we wait for scientific fields to produce reliable evidence that passes muster when severely critical appraisal methods are applied, interventions should be developed based on reasonable theory and outcomes should be monitored to determine effectiveness. Effectiveness of interventions designed using unreliable evidence should never be assumed.

While the main contribution of our Article might seem quite negative, we believe that the future of psychology and its legal applications is bright. Much has improved since Münsterberg's day: speculative claims about the accuracy of diagnostic tools have gradually been replaced by controlled studies. That said, seemingly more sophisticated studies present dangers of their own, especially studies designed to address important societal issues. Although the field of experimental psychology is undergoing a credibility revolution, and we believe that a time will come when the credibility levels of findings will be evident to all, we need an interim approach. Appliers of modern psychological science must adequately acknowledge the field's limitations.

In line with the general credibility movement, we call for the development of a metaresearch agenda for law and psychology. Researchers

---

[307] AUSTRALIAN L. REFORM COMM'N, *supra* note 218, at 111.

are increasingly (1) systematically studying their own methods, (2) developing empirically-informed ways of improving those methods, and (3) testing whether those new practices have the desired effect.[308] Metaresearch studies focusing specifically on experimental psychology have informed the broader movement. We believe that it is time for law and psychology to turn inward.

What would a metaresearch agenda in law and psychology look like? We propose beginning with systematic audits of the existing literature. Audits would measure current research and reporting practices, generally following our review *supra* Part I (which is itself limited in that it mostly relies on examples from the literature rather than a systematic survey of the field). For example, metaresearchers have established protocols for measuring data sharing and reproducibility to ascertain how often studies report sufficient information on employed research methods to allow others to verify their findings.[309] Audits should also measure how often researchers justify their sample sizes and provide transparent effect size calculations.[310] With these data in hand, interventions to improve reporting can be implemented, and appliers of research will have sufficient information to assess the general credibility of particular fields of research and to know when they should provide warnings. Interventions might include guidelines promoting cautions about generalizations, statements explaining why effect sizes may or may not accumulate, and warnings about publication bias. As interventions are implemented, researchers will be able to study whether their intended effects have manifested.

Given the applied nature of law and psychology, the field's metaresearch agenda must consider the research translation efforts of legal scholars, courts, and policy-setting bodies. We urge metaresearchers to build on our meta-study[311] by further systematically studying vulnerabilities in the research to action pipeline. Important and unanswered questions include whether researchers cite studies that have been retracted or contradicted by large-scale replication studies and whether courts and policy-setting bodies put more weight on registered studies with larger sample sizes. These studies also can be repeated in future years as new methods for transparently presenting the strengths and weaknesses to users of research are developed.[312]

Finally, greater collaboration between *metaresearchers*, *legal psychologists*, and *research users* can improve law and psychology. Users of research have valuable knowledge about, for instance, currently

---

[308] For a review, *see* Hardwicke et al., *supra* note 25, at 13.
[309] *See* Tom E. Hardwicke, Joshua D. Wallach, Mallory C. Kidwell, Theiss Bendixen, Sophia Crüwell & John P. A. Ioannidis, *An Empirical Assessment of Transparency and Reproducibility-Related Research Practices in the Social Sciences (2014–2017)*, 7 ROYAL SOC'Y 1, 3-4 (2020).
[310] *See* Chin, *supra* note 116, at 3-4.
[311] *See supra* Part III.
[312] For a paradigm in prevention science, *see* Pamela R. Buckley, Charles R. Ebersole, Christine M. Steeger, Laura E. Michaelson, Karl G. Hill & Frances Gardner, *The Role of Clearinghouses in Promoting Transparent Research: A Methodological Study of Transparency Practices for Preventive Interventions*, 23 PREVENTION SCI. 787, 788 (2022).

understudied contexts and specific effect sizes that are practically relevant to their decisions. Metaresearchers and other methodology specialists can boost the usefulness of their research by communicating and collaborating with users to develop research questions and experiment designs. Finally, as even Wigmore eventually acknowledged,[313] if legal psychologists reform their research practices, they can become integral to bridging the gap between questions of law and research paradigms that help answer those questions.

[313] Magner, *supra* note 5, at 131 ("Thus Wigmore appears to have remained convinced throughout his life of the potential relevance and worth of psychological research into forensic questions").

CREDIT STATEMENT

Conceptualization: Jason M. Chin and Alex O. Holcombe.

Data curation: Jason M. Chin.

Formal analysis: Jason M. Chin.

Funding acquisition: Jason M. Chin, Kathryn Zeiler.

Investigation: Jason M. Chin, Alex O. Holcombe, and Ann Guo.

Methodology: Jason M. Chin.

Project administration: Jason M. Chin.

Supervision: Jason M. Chin and Alex O. Holcombe.

Visualization: Jason M. Chin.

Writing - original draft: Jason M. Chin and Kathryn Zeiler.

Writing - review & editing: Jason M. Chin, Alex O. Holcombe, Patrick S. Forscher, and Kathryn Zeiler.[314]

[314] CRediT statement generated using the 'tenzing' app. *See* Alex O. Holcombe, Marton Kovacs, Frederik Aust & Balazs Aczel, *Documenting Contributions to Scholarly Articles Using CRediT and Tenzing*, 15 PLOS ONE e0244611, 5 (2020).