ARTICLE

CAUSE, EFFECT, AND THE STRUCTURE OF THE SOCIAL WORLD

MEGAN T. STEVENSON[*]

ABSTRACT

This Article is built around a central empirical claim: most reforms and interventions in the criminal legal space are shown to have little lasting effect when evaluated with gold standard methods. While this might be disappointing from the perspective of someone hoping to learn what levers to pull to achieve change, I argue that this teaches us something valuable about the structure of the social world. When it comes to the type of limited-scope interventions that lend themselves to high-quality evaluation, social change is hard to engineer. Stabilizing forces push people back toward the path they would have been on absent the intervention. Cascades—small interventions that lead to large and lasting changes—are rare. And causal processes are complex and context dependent, meaning that a success achieved in one setting may not port well to another.

This has a variety of implications. It suggests that a dominant perspective on social change—one that forms a pervasive background for academic research and policymaking—is at least partially a myth. Understanding this shifts how we should think about social change and raises important questions about the process of knowledge generation.

[*] Thanks for invaluable feedback from Shawn Bushway, Aaron Chalfin, Jason Chin, Eric Chyn, Erin Collins, Scott Cunningam, Brandon Garrett, Ben Grunwald, Thomas Frampton, Barry Friedman, Paul Heaton, Andrew Hayashi, David Hoffman, Dan Hopkins, Alec Karakatsanis, Ben Levin, Sandy Mayson, Aurelie Ouss, Justin Pickett, JJ Prescott, Chris Slobogin, Holger Spamann, and Malcolm Stevenson, as well as numerous participants in the Neighborhood Criminal Justice Roundtable, the USC Law & Economics Colloquium, CELS Toronto, Brooklyn Law School Faculty Workshop, George Mason Law & Economics Colloquium, the UVA Summer Workshop, and the UVA Faculty Workshop. Thanks also to the UVA Law Librarians for their excellent support, particularly Ben Doherty, and to Billi Jo Morningstar for copyediting. Many thanks to Jeremy Brunner and others at the *Boston University Law Review* for their invaluable work editing this Article.

CONTENTS

INTRODUCTION

This Article is built around a central empirical claim: most reforms and interventions in the criminal legal space are shown to have little lasting impact when evaluated with gold-standard methods of causal inference.[1] This claim will not be controversial to anyone immersed in the literature.[2] But, like a dirty secret, it almost never gets seriously acknowledged or discussed. Nor is it widely known beyond the small circle of people trained in statistical methods of causal inference. The research that people hear about shows the rare cases of success; the remainder gets filtered from public view.

The goal of this Article is to establish this claim and discuss its implications. First and foremost, it teaches us something about the structure of the social world. It teaches us that, at least when it comes to the type of limited-scope reforms evaluated by gold standard causal inference methods, change is hard to engineer. To the extent that an intervention temporarily alters some aspect of a person's life, stabilizing forces usually steer them back onto the path they would have been on absent the intervention. Cascades—small interventions that lead to large and lasting impact—are rare. And causal processes are complex and context dependent, meaning that an intervention that happens to succeed in one place may not succeed in another.

That's not to say that the social world is static: to the contrary, it's changing all the time. This is a claim about the ability to *engineer* such change using a particular type of intervention. And it pertains most directly to the type of interventions that get evaluated using randomized controlled trials ("RCTs").

RCTs first gained prominence in the medical space as a way of testing whether a drug, vitamin, or exercise regime delivered on its purported benefits.[3] Hoping to replicate the medical-context success, the social sciences have embraced RCTs.[4] This movement, sometimes called "evidence-based reform," is predicated on the idea that RCTs enable us to identify which reforms and

[1] *See infra* Part II.

[2] Indeed, it has been dubbed the Iron Law of Evaluation. *See* Peter H. Rossi, *The Iron Law of Evaluation and Other Metallic Rules*, *in* 4 RESEARCH IN SOCIAL PROBLEMS AND PUBLIC POLICY: A RESEARCH ANNUAL 3, 4 (Joann L. Miller & Michael Lewis eds., 1987) ("*The Iron Law of Evaluation*: The expected value of any net impact assessment of any large scale social program is zero.").

[3] *See* Arun Bhatt, *Evolution of Clinical Research: A History Before and Beyond James Lind*, 1 PERSPS. CLINICAL RSCH. 6, 8 (2010) (outlining rise of RCTs in medical space beginning in early 20th century).

[4] Cecelia Klingele, *The Promises and Perils of Evidence-Based Corrections*, 91 NOTRE DAME L. REV. 537, 553-54 (2015) ("The evidence-based approach pioneered in medicine quickly translated to other fields requiring clinical judgment, such as nursing and psychology, and then later to the social sciences and other structured fields of inquiry, including education." (footnote omitted)).

interventions are successful, and which are not.[5] The goal is to engineer society to function more effectively.

Part I provides an overview of RCTs, their power, and their limits. Those who are already knowledgeable in this area can skip Sections I.A-B, but I recommend they at least skim Sections I.C-D. These sections explain a primary reason I focus on RCTs (reduced distortion from publication bias) and discuss the types of questions RCTs answer.

Part II provides an overview of fifty-plus years of RCTs in the criminal legal space.[6] The interventions evaluated include things like job training programs, therapy, intensive probation, noncognitive skills training, "swift, certain, and fair" sanctioning, boot camps, housing vouchers, and so forth. Note that the interventions include both "tough on crime" approaches as well as more supportive ones. Most of the interventions evaluated were shown to have little to no lasting effects. And the ones that were successful in an original evaluation usually were not successful when evaluated in other settings.

I argue in Part III that this teaches us something important about the structure of the social world: when it comes to the type of limited scope interventions evaluated by RCTs, the world does not operate in a simple mechanical fashion. Stabilizing forces limit the impact of reforms and interventions. And success in one time and place rarely ports well to another. Part III also discusses the scope of my claim. In its narrowest view, this Article could be read as being about reforms and interventions in the criminal justice space. If that's all that the reader feels is warranted, I have no objection to this narrow interpretation. However, this is not my stance, and I present several arguments for why my claims are not unique to criminal justice.

I discuss implications in Part IV. First, the evidence calls into question a widespread view about the structure of the social world, which I call the *engineer's view*. Under the engineer's view, social processes are structured and manipulable. RCTs and other causal inference methods are used to map the functioning of the machine, to see what impact a particular lever has.[7] They can

[5] *See* Esther Duflo, Speech at the Nobel Prize Banquet (Dec. 10, 2019) (transcript available at https://www.nobelprize.org/prizes/economic-sciences/2019/duflo/speech/ [https://perma.cc/FMT9-QBCF]) ("We believed that like the war on cancer, the war on poverty was not going to be won in one major battle, but in a series of small triumphs, and with no doubt many setbacks along the way. To assess the progress, we adopted the methods of randomized controlled trials, popular in medicine but not really used in economics at the time."); Erin Collins, *Abolishing the Evidence-Based Paradigm*, 48 BYU L. REV. 403, 403 (2022) ("The belief that policies and procedures should be data-driven and 'evidence-based' has become criminal law's leading paradigm for reform.").

[6] I focus on RCTs done "in the field" as opposed to those conducted in laboratory settings.

[7] *See, e.g.*, *Prior Reforms: Criminal Justice Realignment*, CAL. CTS., https://www.courts.ca.gov/75474.htm [https://perma.cc/6FSB-G726] (last visited Nov. 10, 2023) ("Perhaps the most important reform in state sentencing and corrections practice taking place today is the incorporation of principles of evidence-based practice into state sentencing and

be used to identify interventions that yield consistent and replicable success.[8] The uncertainty of reform is minimized because interventions can be piloted before scaling up.[9]

When it comes to the type of limited-scope interventions evaluable via RCT and other quasi-experimental methods, the engineer's view appears to be mostly a myth. More than fifty years of RCT evidence shows the limits in our ability to engineer change with this type of intervention.[10] And when it comes to larger scale or systemic reform, the engineering project is imbued with uncertainty. You can't pilot test systemic reform before scaling up; predictions about its impact depend on heavily contested theories and assumptions.

The engineer's view is widespread among scholars, policymakers, and philanthropic organizations.[11] But its reach extends beyond policy wonks and advocates of evidence-based reform.[12] Narratives that place structure on a social problem and suggest actions to ameliorate it are a common way in which people interpret the world. Many of these narratives embed the engineer's view as part of their folk wisdom.

Recognizing that the world doesn't operate in this fashion opens new doors for thinking about social change.[13] If the type of limited-scope intervention evaluated by RCTs has limited impact, then those who hope to achieve change can either: (1) focus on interventions with immediate and direct benefit; (2) continue with limited-scope interventions in the hopes that they have the type of benefit that is difficult to measure; or (3) seek systemic reform, with all its uncertainties.

Why has the engineer's view become so prevalent if it is not supported by the evidence? Although a full answer is beyond the scope of this Article, I offer a few observations.[14] In part, it remains dominant because people are only exposed

corrections policy and practice. . . . EBP refers to outcome-focused approaches and interventions that have been scientifically tested in controlled studies and proven effective.").

[8] *See, e.g.*, Peter Dizikes, *MIT Economists Esther Duflo and Abhijit Banerjee Win Nobel Prize*, MIT NEWS (Oct. 14, 2019), https://news.mit.edu/2019/esther-duflo-abhijit-banerjee-win-2019-nobel-prize-economics-1014 [https://perma.cc/GN63-AME2] ("[Jameel Poverty Action Lab] also examines which kinds of local interventions have the greatest impact on social problems, and works to implement those programs more broadly, in cooperation with governments and NGOs.").

[9] *See, e.g.*, Nat'l Inst. of Just., *CrimeSolutions: The Evidence-Based Guide for Justice Agencies in Search of Practices and Programs That Really Work*, CORR. TODAY, Nov./Dec. 2021, at 12, 12 ("[J]ustice agencies seek assurance the particular science underlying an existing or contemplated program or practice is sound, and the program or practice, if properly implemented, can work as intended. NIJ has an established, evidence-based online resource to help justice agencies find and refine reliable solutions." (footnote omitted)).

[10] *See infra* Part III.

[11] *See infra* Section IV.A.

[12] *See infra* Section IV.A.

[13] *See infra* Part IV.

[14] *See infra* Section IV.A.

to research that has made it through the distorting filter of research and publication incentives.[15] This filter suppresses research that isn't statistically significant, sufficiently novel, or otherwise exciting.[16] Most people are only aware of the tiny set of studies that made it through the sieve.[17] And these studies are biased toward showing that the intervention evaluated was more successful than it actually was.[18]

I am not the first person to claim that interventions are rarely successful. In the 1970s, criminologist Robert Martinson made the infamous proclamation that "nothing works" when it comes to prisoner rehabilitation, although he later walked back that claim.[19] In the 1980s, sociologist Peter Rossi argued that the failure of social programs was so ubiquitous that it should be known as the *Iron Law of Evaluation*.[20] But such arguments have gone out of fashion. Nowadays, few take such a broad scope view on the patterns of empirical research. And none, as far as I am aware, move from the empirical evidence to a discussion of what it teaches us about the structure of the social world, or about how to achieve social change.

Many of the claims made in this Article derive from personal experience. I've spent my entire career operating within a research paradigm heavily influenced by the engineer's view. My Ph.D. training is in causal inference methods, and I've been doing econometric research on the criminal legal system for the last ten years. The arguments I present in this Article are the results of a slow process

[15] *See* Garret Christensen & Edward Miguel, *Transparency, Reproducibility, and the Credibility of Economics Research*, 56 J. ECON. LITERATURE 920, 928 (2018) ("Taken together, a growing body of evidence indicates that publication bias is widespread in economics and many other scientific fields.").

[16] *See* Annie Franco, Neil Malhotra & Gabor Simonovits, *Publication Bias in the Social Sciences: Unlocking the File Drawer*, 345 SCI. 1502, 1504 (2014) ("[W]e found that some researchers anticipate the rejection of such papers [with null findings] but also that many of them simply lose interest in 'unsuccessful' projects."); Isaiah Andrews & Maximilian Kasy, *Identification of and Correction for Publication Bias*, 109 AM. ECON. REV. 2766, 2767 (2019) ("Estimates based on our replication approach suggest that results significant at the 5 percent level are over 30 times more likely to be published than are insignificant results, providing strong evidence of selectivity.").

[17] *See* Andrews & Kasy, *supra* note 16, at 2766 ("Some empirical results are more likely to be published than others. Selective publication leads to biased estimates and distorted inference.").

[18] *See* John P. A. Ioannidis, T. D. Stanley & Hristos Doucouliagos, *The Power of Bias in Economics Research*, 127 ECON. J. F236, F236 (2017) ("[N]early 80% of the reported effects in these empirical economics literatures are exaggerated; typically, by a factor of two and with one-third inflated by a factor of four or more."); Jason M. Chin & Kathryn Zeiler, *Replicability in Empirical Legal Research*, 17 ANN. REV. L. & SOC. SCI. 239, 242 (2021) ("Irreplicable practices employed in experimental work likely contribute to the surprisingly high number of false and inflated discoveries in the published literature." (citation omitted)).

[19] CARL SIFAKIS, *Martinson, Robert (1927-1979): Sociologist Author of "Nothing Works" Theory*, *in* THE ENCYCLOPEDIA OF AMERICAN PRISONS 157, 157 (2003).

[20] *See* Rossi, *supra* note 2, at 4.

of transformation in how I see the world: one that would not be possible without an insider's perspective on both empirical research and on the incentives empirical researchers face.

I. WHAT IS AN RCT? WHY IS IT POWERFUL?

Randomized controlled trials, or RCTs, are often referred to as the gold standard in empirical research.[21] The vast majority of new drugs approved by the FDA have been evaluated for effectiveness using an RCT.[22] The 2019 Nobel Prize in economics was awarded to researchers known for using RCTs to evaluate programs in development.[23] When it comes to evaluating the effectiveness of an intervention, RCTs are consistently ranked as the highest form of evidence.[24] This section explains what RCTs are and why they are powerful.

A. *What Is an RCT?*

A researcher interested in how intervention X affects outcome Y can run an experiment. In this experiment, the researcher randomly sorts participants into two (or more) groups.[25] Because the participants were randomly sorted, the average attributes of each group should be similar, particularly if the sample size

[21] *See, e.g.*, Eduardo Hariton & Joseph J. Locascio, *Randomised Controlled Trials—The Gold Standard for Effectiveness Research*, 125 BJOG: INT'L J. OBSTETRICS & GYNAECOLOGY 1716, 1716 (2018); Marianne Razavi et al., *US Food and Drug Administration Approvals of Drugs and Devices Based on Nonrandomized Clinical Trials: A Systematic Review and Meta-analysis*, JAMA NETWORK OPEN, Sept. 11, 2019, at 1, 7 ("A system of RCTs is widely considered the most reliable vehicle for advances in therapeutics that result in development of slightly more than 50% of new treatments that are superior to standard treatments.").

[22] *See* Razavi et al., *supra* note 21, at 8 (noting that among 677 drug and medical device applications, only 10% were approved by FDA using non-RCT methods).

[23] Kelsey Piper, *The Nobel Went to Economists Who Changed How We Help the Poor. But Some Critics Oppose Their Big Idea.*, VOX (Dec. 11, 2019, 9:00 AM), https://www.vox.com/future-perfect/2019/12/11/20938915/nobel-prize-economics-banerjee-duflo-kremer-rcts [https://perma.cc/8HRY-AJ4S] ("The Nobel Prize in Economics awarded to Esther Duflo, Abhijit Banerjee, and Michael Kremer . . . was a big win for a scientific approach they've championed: randomized controlled trials . . . .").

[24] *See, e.g.*, *CrimeSolutions Programs by the Numbers*, NAT'L INST. JUST. (Sept. 7, 2022), https://perma.cc/3PRW-SRYW [hereinafter *CrimeSolutions Programs by the Numbers*] ("Most social scientists consider random assignment to lead to the highest level of confidence . . . ."); Patricia B. Burns, Rod J. Rohrich & Kevin C. Chung, *The Levels of Evidence and Their Role in Evidence-Based Medicine*, 128 PLASTIC & RECONSTRUCTIVE SURGERY 305, 306 (2011) (describing RCT's place at top of evidence hierarchy in medical research).

[25] For a general description of RCTs in social science, see JOSHUA D. ANGRIST & JÖRN-STEFFEN PISCHKE, MASTERING 'METRICS: THE PATH FROM CAUSE TO EFFECT ch. 1 (2015).

is large.[26] The researcher then assigns one group to "treatment" and another group to be the "control."[27] In a medical setting, the treatment group might receive an experimental medicine, whereas the control group gets a placebo. In social science research, the treatment group might be placed into substance abuse counseling, while the control group is given a more basic support package. At the end of the trial, researchers compare outcomes across the groups.[28] If the treatment group has lower rates of, say, positive drug tests, the authors might infer that the substance abuse program caused a reduction in drug use.

The process of inferring that intervention X (the substance abuse program) had a causal impact on outcome Y (future positive drug tests) is called *causal inference*.[29] Causal inference in empirical research entails two types of inquiries. One involves establishing that the relationship between X and Y is statistically significant.[30] To properly understand what "statistical significance" means, let's engage in a thought experiment. Consider a substance abuse treatment program for which 200 people applied. 100 of those were randomly selected to participate in the program (the treatment group) and 100 were not admitted (the control group). One year after completion of the program, all 200 people underwent a drug test. If the program was completely ineffective, would we expect the exact same number of positive drug tests for the treatment group and control group? No, this is unlikely. Some amount of random variation is to be expected. If the program was ineffective, you might easily see, say, 22 positive tests in the treatment group and 24 in the control group. You might occasionally see 20 positive tests for the treatment group and 27 positive tests for the control group. Very, very occasionally, you might see 18 positive tests in the treatment group and 29 in the control group.

Statistical significance only tells us that you would rarely see such a large difference in outcomes across the treatment and control groups due to random variation alone.[31] In the previous example, the relationship between the treatment program and future drug tests would not be statistically significant if

[26] *Id.* at 13 ("Two randomly chosen groups, when large enough, are indeed comparable. This fact is due to a powerful statistical property known as the *Law of Large Numbers* (LLN).").

[27] *See id.* at 14.

[28] *See id.* at 21 (describing methods of comparing results).

[29] "Causality" in social science usually refers to the effects of causes, not the causes of effects. *See* Paul W. Holland, *Statistics and Causal Inference*, 81 J. AM. STAT. ASS'N 945, 945 (1986) ("The emphasis here will be on *measuring the effects of causes* because this seems to be a place where statistics, which is concerned with measurement, has contributions to make.").

[30] For an overview of statistical inference, see LEE EPSTEIN & ANDREW D. MARTIN, AN INTRODUCTION TO EMPIRICAL LEGAL RESEARCH ch. 1 (2014).

[31] ANGRIST & PISCHKE, *supra* note 25, at 44 ("Statistically significant results provide strong evidence of a treatment effect, while results that fall short of statistical significance are consistent with the notion that the observed difference in treatment and control means is a chance finding.").

there are 22 positive tests in the treatment group and 24 in the control group. This is because random chance could have easily generated these small differences in outcomes. But if there were 18 positives in treatment and 29 in control, this relationship would be statistically significant.[32] Such a large difference in outcomes is highly unlikely to arise solely from chance.

Once a relationship is said to be statistically significant, the next inquiry involves determining whether the relationship is *causal*. Effectively, this entails ruling out other explanations for the relationship.[33] RCTs are particularly well-suited for this.[34] If you have a sufficiently large group of people in the study, randomization ensures that the treatment and control groups will be identical in expectation, meaning that the attributes across each group are likely to be similar. For example, it would be difficult to argue that the reason there were only 18 positive tests in the treatment group while there were 29 in the control group is because the treatment group had better family support. The treatment group was randomly selected, therefore the level of family support across each group should be similar, as should other characteristics. This helps to rule out a variety of alternative explanations for a relationship.

Note that the second level of inquiry—ruling out other explanations—is not primarily statistical.[35] Certain types of research designs, like RCTs, can provide particularly convincing evidence of causality by virtue of their ability to rule out alternative explanations for a relationship.[36] But causality is established via inferential reasoning, not a formal test.[37]

---

[32] A difference is statistically significant if one would observe a difference that large less than 5% of the time if the "true" probability of a positive test were equal between the two groups. *See* EPSTEIN & MARTIN, *supra* note 30, at 142 (noting 5% as conventional level for assessing statistical significance). This is calculated using formulas that describe the frequency of observed events under an assumed distribution. *See* ANGRIST & PISCHKE, *supra* note 25, at 36-39 (describing sample variance).

[33] More formally, a statistical approach to causal inference assumes that the average outcome of two large, randomly chosen groups would be comparable absent the intervention. Thus, the average outcome of the control group provides an estimate for what the average outcome of the treatment group would have been absent the intervention. The causal effect of treatment is, therefore, the difference between the average observed outcomes for the treatment group and the average observed outcomes for the control group. For a more detailed discussion, see Holland, *supra* note 29, at 946-48.

[34] *See id.* at 946.

[35] Alex Broadbent, Jan P. Vandenbroucke & Neil Pearce, *Response: Formalism or Pluralism? A Reply to Commentaries on 'Causality and Causal Inference in Epidemiology*,*'* 45 INT'L J. EPIDEMIOLOGY 1841, 1849 (2016) (noting that counterfactual contrasts necessary for causality "are not read or calculated from data, but inferred from it").

[36] *See CrimeSolutions Programs by the Numbers*, *supra* note 24.

[37] James J. Heckman, *The Scientific Model of Causality*, 35 SOCIO. METHODOLOGY 1, 2 (2005) [hereinafter Heckman, *The Scientific Model of Causality*] ("[C]ausality is a property of a model of hypotheticals. . . . A model is in the mind. As a consequence, causality is in the mind.").

B. *Selection Bias*

This Article is built around evidence derived from RCTs. Of course, this is not the only type of evidence available pertaining to the causal structure of the social world. A variety of empirical methods attempt to establish causal relationships.[38] I focus on RCT evidence for a few reasons. The first is that RCTs are particularly good at ensuring that the two groups compared—treatment and control—are otherwise similar except for exposure to the intervention.[39] As discussed previously, this otherwise similar condition helps to rule out alternative explanations for a correlation. If the groups aren't randomly assigned, they may differ in meaningful ways. This is formally known as "selection bias" and is a major challenge in estimating causal effects.[40] Imagine I wanted to evaluate the effects of a drug treatment program but was not able to run an RCT. Instead, I simply identified 100 people who had taken the program and 100 people who had not. In order to make them at least somewhat comparable, let's assume that everyone—both those who had taken the program and those who had not—had a prior conviction for drug possession. Let's say one year later, I found that 18 of the program participants had a positive drug test, and 29 of the people who had not taken the program had a positive drug test. As previously, this would be a statistically significant correlation, but would it be *causal*? Who knows! In this scenario, there would likely be a number of differences across the two groups. One group chose to enter the drug treatment program, and had the resources to do so. The other group did not. There are a variety of reasons why the first group might be expected to have fewer positive tests, regardless of the efficacy of the program.

Selection bias is hard to correct for using statistical methods.[41] The researcher can account for differences observed in the data by "controlling" for them in a regression or "matching" the two groups so that those who participated in the drug treatment program look otherwise similar to those that did not.[42] But data is limited. In the real world, there will be important reasons why one group enrolled in the drug treatment program and the other did not, and these reasons will rarely be captured in the data set.[43] You can't control for motivation. You

[38] For an overview of alternative causal inference methods, see SCOTT CUNNINGHAM, CAUSAL INFERENCE: THE MIXTAPE (2021).

[39] *See* ANGRIST & PISCHKE, *supra* note 25, at 15 ("[R]andomly assigned groups should be similar in every way, including in ways that we cannot easily measure or observe.").

[40] *Id.* at 11 ("The principal challenge facing masters of 'metrics is elimination of the selection bias that arises from such unobserved differences.").

[41] *See id.*

[42] Kristofer Bret Bucklen, *Randomized Controlled Trials in Correctional Settings*, CORR. TODAY, Sept./Oct. 2020, at 18, 19 ("While various statistical options exist for establishing a comparison group, the RCT is the strongest design because it best establishes equally comparable groups on all known/measurable and unknown/unmeasurable factors.").

[43] *See* ANGRIST & PISCHKE, *supra* note 25, at 11 ("[W]hen observed differences proliferate, so should our suspicions about unobserved differences.").

can't fully control for access to resources. And without such controls, it's very hard to be confident that any difference in outcomes can be attributed to the intervention rather than these other factors.

C. *Publication Bias*

Selection bias is the primary reason why RCTs are considered the gold standard in empirical evidence.[44] However, there is another reason, which gets less attention, but I believe to be just as important: RCT evidence is less likely to be biased by the distorting incentives of the research process, such as a focus on statistical significance and disproportionate rewards for novelty.

It's notoriously difficult to publish research that isn't statistically significant.[45] This has a large biasing effect on published literature.[46] With standard hypothesis testing methods, there will be one false claim of a relationship between a cause and purported effect for every nineteen times that it fails to find support.[47] Yet these nineteen statistically insignificant results will often never see the light of day.[48] What gets published is the single instance where the spurious causal effect is found. With an infinite supply of research questions, and thousands of scholars looking for interesting research, the literature will be full of false causal claims.[49] Moreover, even if there is a causal relationship between treatment and outcome, if we only observe instances in

[44] ALEJANDRO R. JADAD & MURRAY W. ENKIN, RANDOMIZED CONTROLLED TRIALS: QUESTIONS, ANSWERS, AND MUSINGS 29 (2007) ("The main appeal of the randomized controlled trial (RCT) in health care comes from its potential to reduce selection bias. Randomization, if done properly, can keep study groups as similar as possible at the outset, so that the investigators can isolate and quantify the effect of the interventions they are studying.").

[45] *See* Andrews & Kasy, *supra* note 16, at 2767 ("Estimates based on our replication approach suggest that results significant at the 5 percent level are over 30 times more likely to be published than are insignificant results, providing strong evidence of selectivity.").

[46] Christensen & Miguel, *supra* note 15, at 920 ("[T]here is growing evidence documenting the prevalence of publication bias in economics and other scientific fields, as well as specification searching, and widespread inability to replicate empirical findings.").

[47] Roger B. Davis & Kenneth J. Mukamal, *Hypothesis Testing: Means*, 114 CIRCULATION 1078, 1078 (2006) ("[For hypothesis testing] [t]he standard value chosen for level of significance is 5% . . . . This standard means that even if no association between predictor and outcome exists in the population, the investigator is willing to accept a 1 in 20 chance of a false-positive conclusion that an association does exist.").

[48] *See* Andrews & Kasy, *supra* note 16, at 2767.

[49] Even if the causal claim is correct, a focus on statistical significance will inflate its magnitude. Statistical significance is determined by a ratio between the estimated effect size and the amount of noise. The result will be statistically significant only when the estimated effect size is larger. *See* ANGRIST & PISCHKE, *supra* note 25, at 21 ("Differences that are larger than about two standard errors are said to be statistically significant . . . .").

which the estimated effect is statistically significant, the magnitude of the effect will be inflated.[50]

In an ideal world, such false causal claims would be overturned by subsequent research. Unfortunately, this subsequent research rarely happens.[51] The first scholar to demonstrate a causal relationship between X and Y can reap large professional benefits: prestigious publications, respect from peers, increased likelihood of tenure, etc.[52] But, in a field that rewards novelty, attempts to replicate the original result have little upside for the researcher's career. If a subsequent study confirms the initial result, it will generally publish in a journal of lower prestige, if at all.[53] If a subsequent study yields opposite results, it may earn the resentment of the original scholar whose work was challenged.[54] Young scholars are often advised to avoid attempting to replicate the work of others.[55]

Distorting incentives such as these affect every stage of the research production process.[56] Researchers must choose which questions to ask, whether to continue a study after some preliminary research has been done, which outcomes and specifications to report, whether to write up a project, and whether

[50] This is particularly true when studies are underpowered, meaning the sample size is too small to be able to reliably identify the causal relationship. *See* Andrew Gelman & John Carlin, *Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors*, 9 PERSPS. PSYCH. SCI. 641, 641, 647 (2014) ("In noisy, small-sample settings, statistically significant results can often be misleading. . . . Using statistical significance as a screener can lead researchers to drastically overestimate the magnitude of an effect . . . .").

[51] *See* Sander L. Koole & Daniël Lakens, *Rewarding Replications: A Sure and Simple Way To Improve Psychological Science*, 7 PERSPS. PSYCH. SCI. 608, 609-10 (2012) (outlining neglect of replication research in psychological science literature); Chin & Zeiler, *supra* note 18, at 242 (describing lack of replication in empirical legal research).

[52] *See* Sarah Necker, *Scientific Misbehavior in Economics*, 43 RSCH. POL'Y 1747, 1747 (2014) ("Science has been compared to a winner-take-all market in which rewards are only granted to those first to make a discovery and therefore obtain recognition from peers." (citation omitted)).

[53] *See* William H.J. Hubbard, *A Replication Study Worth Replicating: A Comment on Salmanowitz and Spamann*, 58 INT'L REV. L. & ECON. 1, 1 (2019) ("[I]f the original study successfully replicates, then the replication study is branded as something worse than wrong: it is *uninteresting*.").

[54] *Id.* ("[I]f the original study fails to replicate, the authors of the original study may feel attacked—and may retaliate, hurting the replication study's author's chances for professional advancement . . . .").

[55] *See id.*

[56] For example, funding incentives can also create pressure to show that the intervention was effective, often because the funder has already invested resources or reputation in the intervention. Angus Deaton, *Randomization in the Tropics Revisited: A Theme and Eleven Variations* 12-13 (Nat'l Bureau of Econ. Rsch., Working Paper No. 27600, 2020), http://www.nber.org/papers/w27600.pdf [https://perma.cc/5UFM-LQYY] [hereinafter Deaton, *Randomization in the Tropics Revisited*] ("Funders who have spent large sums on an RCT often exert pressure to find at least one subgroup for which the treatment was effective.").

to seek publication.[57] Once the researcher has done all they can do, there is a separate question of whether the paper will be published, read, discussed, and cited. The research that most people are aware of is the tiny sliver that makes it through this filtration process.

This process is particularly pernicious when there is a large number of "researcher degrees of freedom," or the ability to tailor results to get the desired effect.[58] When researchers have a variety of choices they can make in their study, these choices are almost inevitably made with incentives in mind.[59] In one recent anonymous survey of quantitative criminologists, 39% reported having changed the analysis after an earlier one wasn't statistically significant and 43% said they failed to report null results.[60] In an anonymous survey of European economists, roughly one-third responded yes each to having: searched for control variables until they found the desired result, presented results selectively so that they confirmed one's argument, or engaged in some other "p-hacking"[61] method.[62] In other anonymous surveys, some 2-5% of social scientists reported having

[57] *See* Franco et al., *supra* note 16, at 1504 ("[W]e found that some researchers anticipate the rejection of such papers [with null findings] but also that many of them simply lose interest in 'unsuccessful' projects."); Christensen & Miguel, *supra* note 15, at 922 ("[W]e might hope that the robustness checks typically demanded of scholars in seminar presentations and during journal peer review manage to keep the most extreme forms of bias in check. Yet we believe most economists would agree that there remains considerable wiggle room in the presentation of results in practice, in most cases due to behaviors that fall far short of outright fraud.").

[58] *See* Nick Huntington-Klein et al., *The Influence of Hidden Researcher Decisions in Applied Microeconomics*, 59 ECON. INQUIRY 944, 945-47 (2021) (describing how researcher degrees of freedom, even when applied in good faith, can result in support of almost any hypothesis).

[59] Eva Vivalt, *Specification Searching and Significance Inflation Across Time, Methods and Disciplines*, 81 OXFORD BULL. ECON. & STAT. 797, 799 (2019) ("Specification searching [any process that leads to statistical significance being inflated] is intimately related to publication bias, as researchers may engage in specification searching in anticipation of journals selectively accepting papers with significant results . . . .").

[60] Jason M. Chin, Justin T. Pickett, Simine Vazire & Alex O. Holcombe, *Questionable Research Practices and Open Science in Quantitative Criminology*, 39 J. QUANTITATIVE CRIMINOLOGY 21, 31 tbl.3 (2023).

[61] Abel Brodeur, Nikolai Cook & Anthony Heyes, *Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics*, 110 AM. ECON. REV. 3634, 3634-35 (2020) ("The term *p*-hacking refers to a variety of practices that a researcher might (consciously or unconsciously) use to generate 'better' *p*-values, perhaps (but not necessarily) in response to the difficulty of publishing statistically insignificant results." (citations omitted)); Vivalt, *supra* note 59, at 798 ("[O]ne way in which p-hacking may occur is by authors including different combinations of control variables until finding a significant result . . . .").

[62] *See* Necker, *supra* note 52, at 1751 tbl.2 (reporting: 32.18% of respondents "presented empirical findings selectively so that they confirm one's argument"; 36.49% "searched for control variables until you got the desired results"; and 37.93% "stopped statistical analysis when you had a desired result").

gone so far as to falsify data.[63] And these are almost certainly lower bounds on the extent of this type of behavior.

All empirical research is subject to distortion by incentives.[64] But some research designs are more susceptible than others.[65] RCTs have fairly stringent requirements in terms of the elements needed.[66] This reduces the researcher degrees of freedom and therefore reduces the latitude to tweak the analysis to get the desired results.[67] Given their strong reputation as the gold standard in research, RCTs are also easier to publish than other research designs, even if the results aren't statistically significant.[68] They are expensive and time-consuming,

[63] *See* John A. List, Charles D. Bailey, Patricia J. Euzent & Thomas L. Martin, *Academic Economists Behaving Badly? A Survey on Three Areas of Unethical Behavior*, 39 ECON. INQUIRY 162, 165, 167 (2001) (noting 4.2-4.5% of economists report having falsified research data and economists speculate that 5.1-7.0% of research in top thirty journals is falsified); Daniele Fanelli, *How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data*, PLOS ONE, May 29, 2009, at 1, 10 (noting meta-analysis of eighteen surveys of academic misbehavior found 2% reporting data fabrication and 34% admitting to lesser forms of academic misconduct).

[64] *See* Christensen & Miguel, *supra* note 15, at 922 ("[W]e might hope that the robustness checks typically demanded of scholars in seminar presentations and during journal peer review manage to keep the most extreme forms of bias in check. Yet we believe most economists would agree that there remains considerable wiggle room in the presentation of results in practice, in most cases due to behaviors that fall far short of outright fraud.").

[65] *See* Chin & Zeiler, *supra* note 18, at 242 ("[G]reater analytic flexibility—enabled by a lack of replicability—leads to less trustworthy results.").

[66] Bonnie Sibbald & Martin Roland, *Understanding Controlled Trials: Why Are Randomised Controlled Trials Important?*, 316 BRITISH MED. J. 201, 201 (1998) (outlining rigorous RCT requirements). There has been a movement to further reduce researchers' abilities to manipulate results in RCTs by preregistering the research design (i.e., declaring outcomes of interest, statistical specification, etc., before the data has been collected). However, fidelity to the preregistered design varies. *See* Alese Wooditch, Lincoln B. Sloas, Xiaoyun Wu & Aleisha Key, *Outcome Reporting Bias in Randomized Experiments on Substance Use Disorders*, 36 J. QUANTITATIVE CRIMINOLOGY 273, 274 (2020) ("For instance, it has been found in medical research that deviations from the pre-specified protocol are common and that these deviations are associated with higher observed effect sizes." (citation omitted)). When authors deviate from the preregistered plan, estimates are substantially larger, consistent with cherry picking. *Id.*

[67] Brodeur et al., *supra* note 61, at 3636 ("[O]ur results suggest that the [instrumental variables] and, to a lesser extent, [differences-in-differences] research bodies have substantially more *p*-hacking and/or selective publication than those based on RCT and [Regression Discontinuity Design]. . . . [One] potential explanation is that some methods offer researchers different degrees of freedom than others. For instance, when using a non-experimental method like [instrumental variables] there are many points at which a researcher exercises discretion in ways that could affect statistical significance.").

[68] Vivalt, *supra* note 59, at 810 ("[W]e might expect that RCTs would exhibit fewer traces of bias due to their being more likely to be published independent of their results and due to their increased rigor perhaps making specification searching more difficult."); Brodeur et al.,

so once someone has put in the effort to conduct one, they are less likely to abandon it.[69] In contrast, if someone begins a project using a less reputable research design, they often won't even bother to write it up if results are not statistically significant.[70]

For all these reasons, published RCT studies are less likely to be biased by researcher incentives than other types of causal inference research. And any bias that remains goes in a predictable direction: toward inflating results and over-reporting false causal claims.

D. *The Types of Questions RCTs Can Answer*

RCTs can answer certain types of questions very well, and others not at all. Because this Article's arguments are built on RCT evidence, it's important to be precise about the domain of that evidence: questions answered and answerable by these methods. In particular, RCTs provide evidence about the *causal impact* of *interventions* on *outcomes* for a particular *sample*.[71] Let's break that down.

The term "causality" can be used in a variety of ways.[72] In empirical causal inference research causality refers to the idea that if some external force (e.g., the researcher) changes someone's exposure to an intervention (e.g., a drug treatment program) one would expect to see a change in outcomes (e.g., a future positive drug test).[73] If so, one would say that the drug treatment program had a *causal impact* on the likelihood of testing positive in the future.

The *interventions* evaluated by RCTs need to be manipulable, meaning that they need to be something that can be randomly assigned in the context of an experiment.[74] In practice, this entails a few things. First, interventions need to be *isolable*, meaning the interventions can be separated from other aspects of

supra note 61, at 3636 ("Another potential explanation for our main observations is that the attitudes of editors and/or referees toward null results vary systematically with method. For example, there may be more tolerance of a null result if it is the result of an RCT.").

[69] *See* Franco et al., *supra* note 57, at 1503 ("The baseline probability of publishing experimental findings based on representative samples is likely higher than that of observational studies using 'off-the-shelf' data sets or experiments conducted on convenient samples in which there is lower 'sunk cost' involved in obtaining the data.").

[70] *See* Brodeur et al., *supra* note 61, at 3635-36 (noting absence of statistically insignificant results of published studies using less randomized methods than RCT).

[71] Hariton & Locascio, *supra* note 21, at 1716.

[72] Margaret Mooney Marini & Burton Singer, *Causality in the Social Sciences*, 18 SOCIO. METHODOLOGY 347, 347 (1988) (describing imprecision and changes in causal terminology over time).

[73] This is sometimes referred to as the Neyman-Rubin model. A description can be found in Holland, *supra* note 29, at 946 or Heckman, *The Scientific Model of Causality*, *supra* note 37, at 35 ("Many statisticians and social scientists invoke a model of counterfactuals and causality attributed to Donald Rubin by Paul Holland (1986) but which actually dates back to Neyman (1923).").

[74] *See* Holland, *supra* note 29, at 959.

experience.[75] Socioeconomic class, for instance, is not isolable. You can't randomly assign one group to a higher socioeconomic class because socioeconomic class—a complicated combination of wealth, education, race, career, etc.—is too intertwined in the human experience. Income, on the other hand, is isolable. For instance, one could conduct an RCT which gives a randomly selected group of people a bunch of money. Many such RCTs have been conducted.[76]

Interventions also tend to be of *limited scope*. In RCTs, experiments are likely to be considered unethical if they make an individual appreciably worse off.[77] Cost and other practical constraints also often limit how much better off you can make them.[78] This issue extends beyond RCTs and into the realm of natural experiments and other quasi-experimental research designs.[79] The more you can compare two groups of people that are otherwise similar except for the intervention—as in RCTs—the more you can rule out alternative explanations. But in the real world, it is rare to find groups that vary massively along one dimension but are otherwise statistically identical. If you want to evaluate larger interventions—laws, large-scale social policies, etc.—you are unlikely to find two groups that are truly identical except for that intervention.

RCTs also inevitably focus on the set of *outcomes* selected by researchers as interesting and worthy of study.[80] Researchers generally aren't interested in studying questions where the answer seems obvious.[81] Take a job training program, for instance. If you enroll someone in a job training program that provides subsidized employment, the employer gets discounted labor and the participant gets a job. It's almost mechanical that employment levels will

[75] ANGRIST & PISCHKE, *supra* note 25, at 50.

[76] *See, e.g.*, Gary Burtless, *The Work Response to a Guaranteed Income: A Survey of Experimental Evidence*, *in* LESSONS FROM THE INCOME MAINTENANCE EXPERIMENTS, at 22, 35 (Alicia H. Munnell ed., 1986).

[77] Angus Deaton & Nancy Cartwright, *Understanding and Misunderstanding Randomized Controlled Trials*, 210 SOC. SCI. & MED. 2, 7 (2018).

[78] *See* Sibbald & Roland, *supra* note 66, at 201 (describing cost, ethical, and practical limitations of RCTs in medical science).

[79] *See infra* Section III.B.

[80] An additional constraint on outcomes is that they must be measurable. You cannot directly measure the impact of an intervention on intangible concepts such as well-being or capacity. Such concepts are unmeasurable in a very basic sort of way: there is no agreed-upon metric of well-being or capacity, let alone universal agreement about what such terms mean. However, they would be very thin concepts indeed if they had no relationship to measurable aspects of the lived experience. Generally, well-being is thought to increase with things like health, education, and economic opportunity. These metrics can thus be indirect proxies for well-being.

[81] *See supra* notes 51-63 and accompanying text.

increase while in that program.[82] Social scientists rarely set out to demonstrate the direct, mechanical effect of an intervention.[83] When it is considered, the main goal is usually to measure the *degree* of impact, not to demonstrate that an impact exists.

Instead, researchers tend to be interested in longer-term or indirect effects of a program.[84] They are interested in whether the employment gains persist after the period of subsidization ends,[85] if the job-training program reduces crime, and so forth. Such effects are more speculative, and therefore more in need of evaluation.[86]

Finally, RCTs provide evidence about the causal impact of an intervention on a particular *sample*, meaning the group of people in the study. And it is not always clear whether evidence from one particular sample provides useful information about other times, places, and groups of people.[87] If not, the result lacks "external validity," meaning that information derived from that study does not generalize to other contexts.[88]

A number of commenters have argued that RCTs hold an overly elevated position in the hierarchy of methods.[89] Some highlight the implementation

[82] *See* DAVID BUTLER ET AL., U.S. DEP'T OF HEALTH & HUM. SERVS., OPRE REPORT 2012-08, WHAT STRATEGIES WORK FOR THE HARD-TO-EMPLOY? 64 (2012) ("[A]ll of [the job training programs] increased employment and earnings early in the follow-up period, when participants were in temporary (subsidized) transitional jobs.").

[83] Rossi, *supra* note 2, at 5 ("*The Zinc Law of Evaluation*: Only those programs that are likely to fail are evaluated. . . . It also implies that if a social program is effective, that characteristic is obvious enough and hence policy makers and others who sponsor and fund evaluations decide against evaluation.").

[84] *See, e.g.*, Klingele, *supra* note 4, at 539-41 (describing long-term recidivism goals of those looking at sentencing analyses).

[85] *See, e.g.*, BUTLER ET AL., *supra* note 82, at ES-1 ("[The] programs had a variety of goals, but they all aimed, directly or indirectly, to increase employment and earnings, and most aimed to reduce reliance on public assistance.").

[86] They are also often seen as the primary goal of an intervention. If a job training program did nothing but increase employment during the period of wage subsidization, it might not be seen as cost effective.

[87] Heckman, *The Scientific Model of Causality*, *supra* note 37, at 8.

[88] *Id.*

[89] *See, e.g.*, Deaton & Cartwright, *supra* note 77, at 2 (arguing special status for RCTs is unwarranted); Judea Pearl, *Challenging the Hegemony of Randomized Controlled Trials: A Commentary on Deaton and Cartwright*, 210 SOC. SCI. & MED. 60, 60-61 (2018) (concurring with Deaton & Cartwright and advocating for more theory-driven approach); James J. Heckman, *Randomization and Social Policy Evaluation Revisited* 5-8 (Nat'l Bureau of Econ. Rsch., Technical Working Paper No. 107, 2020), https://www.nber.org/system/files/working_papers/t0107/t0107.pdf [https://perma.cc/5J7N-REPK] [hereinafter Heckman, *Randomization and Social Policy*] (highlighting assumptions that still need to be made with RCTs and various reasons why RCTs might not answer important questions); Robert J. Sampson, *Gold Standard Myths: Observations on the Experimental Turn in Quantitative Criminology*, 26 J.

challenges that can undermine the results.[90] Others point out limits in the type of questions RCTs answer: for instance, that the shift toward using RCTs to evaluate social policies has meant prioritizing small-bore questions over more important (and more difficult to answer) ones.[91] Another objection is that RCT evidence may not generalize well.[92] RCTs are usually conducted as pilot programs with relatively small samples.[93] Lots of resources are devoted to ensuring fidelity of implementation and seeking out success.[94] As a program expands, implementation can get sloppier.[95] It might be harder to find and train qualified individuals to implement the program.[96]

QUANTITATIVE CRIMINOLOGY 489, 490 (2010) (critiquing supposed "gold standard" status ascribed to RCT models and arguing RCTs should be used along with observational approaches).

[90] *See, e.g.*, Richard A. Berk, *Randomized Experiments as the Bronze Standard*, 1 J. EXPERIMENTAL CRIMINOLOGY 417, 429-30 (2005).

[91] *See, e.g.*, Angus Deaton, *Instruments, Randomization, and Learning About Development*, 48 J. ECON. LITERATURE 424, 426 (2010) ("The price for [the success of RCTs] is a focus that is too narrow and too local to tell us 'what works' in development, to design policy, or to advance scientific knowledge about development processes."); Heckman, *Randomization and Social Policy*, *supra* note 89, at 5 (noting focus on RCTs is "part and parcel of a professional obsession in the field of economics to obtain 'causal effects,' even if the effects being identified are without social significance and/or economic meaning"); Sampson, *supra* note 89, at 492 ("[C]riminologists are often concerned with causal processes that take on historical and institutional dimensions that range over long periods of time (sometimes decades) and that are not amenable to randomization.").

[92] Deaton, *Randomization in the Tropics Revisited*, *supra* note 56, at 9 ("A study can be outstandingly done and unassailable, which tells us nothing about whether it generalizes to other settings.").

[93] *See* Heckman, *Randomization and Social Policy*, *supra* note 89, at 26 (discussing how RCTs are mostly implemented on "pilot projects" or "demonstration projects").

[94] *See* Abhijit V. Banerjee & Esther Duflo, *The Experimental Approach to Development Economics*, 1 ANN. REV. ECON. 151, 164 (2009) (noting governmental partners for RCTs tend to have greater "competence and a willingness to implement projects as planned," characteristics which may get "lost when the project scales up").

[95] *See, e.g.*, Monica P. Bhatt, Jonathan Guryan, Jens Ludwig & Anuj K. Shah, *Scope Challenges to Social Impact* 8-10 (Nat'l Bureau of Econ. Rsch., Working Paper No. 28406, 2021), https://www.nber.org/system/files/working_papers/w28406/w28406.pdf [https://perma.cc/3G2N-GX3Z] (explaining potential difficulty of recruiting more skilled counselors as Chicago's Becoming a Man program expanded).

[96] *Id.* Scaling up an intervention can also change the nature of the intervention. Consider an RCT that randomly selects low-income individuals to receive a housing voucher allowing them to live in a high-income neighborhood. *See, e.g.*, Raj Chetty, Nathaniel Hendren & Lawrence F. Katz, *The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment*, 106 AM. ECON. REV. 855, 855 (2016). It is not clear what scaling up such an intervention would entail. If you moved enough low-income people to a high-income neighborhood, that neighborhood would change. And the impact of living in that neighborhood would likely change, too.

These are important limitations, and valid reasons why social science research cannot solely consist of RCTs.[97] But even if the domain of questions answerable by RCTs is limited, it's still an important and interesting domain. The interventions evaluated by RCTs may be of limited scope, but they are *actionable* interventions. They are things that you (or an agency, nonprofit, or government branch) can *do*. You may not be able to directly impact someone's socioeconomic class. But, you can give them an income transfer. You can enroll them in a job training class. You can give them a scholarship to college. The questions answered by RCTs are directly relevant to the activist, NGO, or policymaker who wants to know what they can do to change the world.

Moreover, an intervention that is limited in scope may not be limited in effect. A low-cost intervention targeting a pivotal moment in people's lives could, at least in theory, have large and lasting impact.[98] If change is cumulative, then small interventions can grow in influence over time.[99] Such low-cost, high-benefit interventions are the holy grail of social engineers.

As for the generalizability concerns, similar issues apply to all empirical research.[100] All research focuses on one particular group of people in one particular time and place. An inferential leap must be made every time results from one setting are used to inform expectations in another setting.

This Article is built on an inferential argument. I am arguing that we can draw inferences from more than fifty years of RCTs in the criminal justice space to learn something more broadly about the social world. But first, let me present the evidence.

II. FIFTY-PLUS YEARS OF RCT EVIDENCE

This Part presents the central empirical claim of this Article: that most reforms and interventions in the criminal legal space have little to no lasting effect when evaluated by RCTs, and the occasional success usually fails to replicate when evaluated in other settings.

---

[97] There are a variety of limitations to RCTs, including challenges with implementation, subject attrition, concerns that randomization itself could influence the treatment effect (randomization bias) or otherwise make results less generalizable. *See* Berk, *supra* note 90, at 429-30 ("[R]andomized experiments are not the gold standard. But if the truth be told, there is no gold standard. There can be settings in which the strengths and weaknesses of potential research designs favor an alternative to randomized experiments.").

[98] *See* Bhatt et al., *supra* note 95, at 4, 6 (discussing ways comparatively small interventions could have large impact on given individual).

[99] For one example of such an argument, see Sara B. Heller, *Summer Jobs Reduce Violence Among Disadvantaged Youth*, 346 SCI. 1219, 1219 (2014) [hereinafter Heller, *Summer Jobs Reduce Violence*] ("Offering summer employment at this key point in the life course could make crime a relatively less attractive option, strengthen social bonds, and develop 'soft' skills such as self-efficacy and impulse control.").

[100] Banerjee & Duflo, *supra* note 94, at 161-63 (comparing generalizability issues in RCTs to generalizability issues in observational studies).

A claim about the general tendency of a literature is both hard to prove and hard to refute. It is hard to prove because of its breadth. There exists no single repository that contains all RCTs in the criminal legal space; there is no easy way to catalog and summarize the literature. And it is hard to refute because it is not absolute. There may be occasional instances in which a meaningful, lasting, and replicable causal effect is demonstrated, I simply claim they are rare.

My strategy here is threefold. First, I take a wide lens, and discuss findings from a broad survey study of RCTs in a variety of criminal justice topics. Second, I zoom in on several of the most prominent and influential studies of the last few decades, studies in which the effects were so promising that multiple replication studies were attempted. Third, I move through a variety of popular, highly-studied interventions in criminal justice and discuss the evidence associated with each. Because highly studied interventions are those which researchers and/or funders expect to be most effective, this is a conservative approach, meaning I am selecting a group of topics in which my thesis is less likely to be true. This three-part strategy certainly falls short of definitive proof; those who arrive skeptical of my claim may not walk away fully convinced. Nonetheless, I hope it is eye opening.

In 2006, two criminologists published a survey article of every RCT over the previous fifty years in which: (1) there were at least 100 participants, (2) the study included a measure of offending as an outcome, and (3) the study was written in English.[101] The authors uncovered 122 studies, evaluating interventions such as:

- Counseling/therapy programs;[102]
- Criminal legal supervision, including intensive probation;[103]
- Scared-straight programs;[104]
- Work/job-training programs;[105]
- Drug testing, substance abuse counseling, and drug court;[106]
- Juvenile diversion;[107]
- Policing "hot spots";[108] and
- Boot camps.[109]

Note that these interventions include those associated with a tough-on-crime framework (e.g., scared-straight programs and boot camps) as well as those that provide support and resources (e.g., work/job training programs and

[101] David P. Farrington & Brandon C. Welsh, *A Half Century of Randomized Experiments on Crime and Justice*, 34 CRIME & JUST. 55, 60-61 (2006).

[102] *Id.* at 68, 70-71, 74-75, 92-93, 99, 101, 109.

[103] *Id.* at 68, 79, 81, 107-08.

[104] *Id.* at 98.

[105] *Id.* at 71, 75, 78, 82-83, 93.

[106] *Id.* at 99, 101-02, 107.

[107] *Id.* at 68.

[108] *Id.* at 88-89.

[109] *Id.* at 98.

counseling). Note further that inclusion in this analysis required that the study was written up and disseminated so it could be discovered by the survey authors—a filter that is likely to have eliminated many of the nonstatistically significant results already.[110] Nonetheless, only 29 of the 122 studies (24%) found statistically significant impacts in the desired direction.[111] Furthermore, the estimated treatment effect was in the desired direction in only 77 out of 122 studies (63%).[112] This is better than a coin flip, but not much.

An aside for those with statistical training: some of the studies discussed in this Section are likely underpowered. On an individual level, a null finding would not be informative on an underpowered study.[113] In aggregate, however, a large number of null results suggest that the interventions generally had small or nonexistent effects.[114]

What about the twenty-nine studies with statistically significant impacts? Were these successes just byproducts of publication bias or specification search, or were they interventions with consistent and replicable benefits? Unfortunately, most of the studies were one offs, meaning that there were no attempts to replicate the results in other settings. In fact, across the fifty years surveyed, the authors identified only a few instances in which a successful original study had multiple replication attempts.[115] One was the Minneapolis Domestic Violence Experiment: an RCT in which officers were randomly assigned to either immediately arrest the accused perpetrator after a domestic violence call or to give them advice and order them to separate from their

[110] *See* Franco et al., *supra* note 16, at 1504 ("[W]hat is perhaps most striking . . . is not that so few null results are published, but that so many of them are never even written up . . . .").

[111] Farrington & Welsh, *supra* note 101, at 111 (reviewing two sets of experiments from different time periods—in first set "only nine [of thirty-seven] experiments produced significantly desirable results," and in second set "twenty experiments [out of eighty-five] produced significantly desirable results," for total of 29 out of 122).

[112] *Id.*

[113] Miguel A. Vadillo, Emmanouil Konstantinidis & David R. Shanks, *Underpowered Samples, False Negatives, and Unconscious Learning*, 23 PSYCHONOMIC BULL. & REV. 87, 88 (2016) ("Even though individual underpowered studies may fail to reject the null hypothesis, meta-analysis across a set of such studies may permit modest but real effects to be detected.").

[114] Assume that studies were powered to detect a positive effect (of a minimally policy-relevant size) at least 40% of the time. These would be highly underpowered studies, given that the standard recommended power is 80%. *See id.* at 367. Even given such lack of power, one would expect at least 40% of the studies to show a statistically significant positive effect if such an effect existed.

[115] Beyond the two discussed in this text, Farrington and Welsh also note that there was a failure to replicate the original success of a program giving prisoners special casework attention from welfare officers, and a failure to replicate the original success of a reentry program connecting visitors with financial aid and job training. Farrington & Welsh, *supra* note 101, at 85.

accuser.[116] The study found that an initial arrest led to substantially fewer repeat incidents over the subsequent six months.[117] Because incarceration times were too short for incapacitation to explain the effect, this study seemed to be documenting a scared-straight type phenomenon, in which the experience of arrest changed the arrestee's willingness to engage in future violence.[118]

This study was enormously influential and was used to support mandatory arrest laws across the nation.[119] Its success prompted the National Institute of Justice to fund six subsequent randomized evaluations to see whether the effects were replicated in other settings.[120] Only one of these six found statistically significant benefits of arrest.[121] Combining data across all seven studies, researchers later concluded that arrest did not, in fact, have a consistent or large effect on recidivism.[122]

A second replicated intervention was Multisystemic Therapy ("MST"), a program that aims to reduce juvenile offending by providing therapy and resources to youths, their families, and their peer groups.[123] The original study and three follow-up studies (all RCTs) showed large reductions in recidivism for

[116] Lawrence W. Sherman & Richard A. Berk, *The Specific Deterrent Effects of Arrest for Domestic Assault*, 49 AM. SOCIO. REV. 261, 262 (1984) (randomizing whether misdemeanor domestic assault offenders were arrested, ordered to leave, or given advice or mediation to measure deterrent effect); Joshua D. Angrist, *Instrumental Variables Methods in Experimental Criminological Research: What, Why and How*, 2 J. EXPERIMENTAL CRIMINOLOGY 23, 25-27 (2006) (describing results of Minneapolis Domestic Violence Experiment).

[117] *See* Sherman & Berk, *supra* note 116, at 268 (comparing recidivism rates among offenders who had received varying interventions).

[118] The actual Scared Straight Program, in which youths were brought to adult prisons in order to scare them away from crime, was actually found to increase criminal activity. Anthony Petrosino, Carolyn Turpin Petrosino & John Buehler, *"Scared Straight" and Other Juvenile Awareness Programs for Preventing Juvenile Delinquency: A Systematic Review of the Randomized Experimental Evidence*, 589 ANNALS AM. ACAD. POL. & SOC. SCI. 41, 58 (2003) (finding that randomized trials studied provided "empirical evidence—under experimental conditions—that these programs likely increase the odds that children exposed to them will commit another delinquent offense").

[119] *See* Farrington & Welsh, *supra* note 101, at 86-87 (noting that Department of Justice encouraged police agencies to adopt mandatory arrest laws).

[120] *Id.* at 87.

[121] *See id.*

[122] Janell D. Schmidt & Lawrence W. Sherman, *Does Arrest Deter Domestic Violence?*, 36 AM. BEHAV. SCIENTIST 601, 603-04 (1993) (finding that studies in different cities had inconsistent results on whether arrest reduced recidivism); CHRISTOPHER D. MAXWELL, JOEL H. GARNER & JEFFREY A. FAGAN, NAT'L INST. JUST., THE EFFECTS OF ARREST ON INTIMATE PARTNER VIOLENCE: NEW EVIDENCE FROM THE SPOUSE ASSAULT REPLICATION PROGRAM 1 (2001) ("[R]ather than providing results that were consistent with [the Minnesota Domestic Violence Experiment], the published results from the five replication experiments produced inconsistent findings about whether arrest deters intimate partner violence.").

[123] Farrington & Welsh, *supra* note 101, at 94.

youths assigned to MST.[124] However, all four of these studies were carried out by the man who originated the treatment, raising conflict of interest questions.[125] A more recent meta-analysis analyzed fifteen RCTs and found that the average effect was statistically significant but small.[126] However, the authors also found substantial evidence of publication bias, meaning that studies with statistically insignificant or negative effects were not being published.[127] Once they adjusted for publication bias, MST's effect size was close to zero and not statistically significant.[128]

Multisite replication attempts using RCTs continue to be rare. Another recent example pertains to "swift, certain, and fair" sanctioning as developed in Project Hawaii Opportunity Probation with Enforcement ("Project HOPE").[129] "Swift, certain, and fair" refers to a model in which a probation violation (e.g., a positive drug test) results in a certain and immediate, but relatively mild sanction, such as twenty-four hours in jail.[130] This replaces a more indeterminate system, in which an individual can accrue multiple violations before a judge decides to revoke probation. In such a system, probation revocation often leads to many months of incarceration. An RCT showed that Project HOPE led to large reductions in both drug use and time incarcerated, with long-lasting effects.[131]

This study was exciting to many, not only because of its impressive effects, but also because it supported a set of theories that were gaining popularity at that time.[132] A central theme in behavioral economics is that people are myopic, meaning that they are expected to respond more to the threat of short sentences that would go into effect immediately than to long sentences that might go into

[124] *Id.*

[125] *See id.*

[126] Trudy van der Stouwe, Jessica J. Asscher, Geert Jan J. M. Stams, Maja Deković & Peter H. van der Laan, *The Effectiveness of Multisystemic Therapy (MST): A Meta-Analysis*, 34 CLINICAL PSYCH. REV. 468, 468 (2014) (noting that impacts on delinquency, the primary outcome, were small but significant).

[127] *Id.*

[128] *See id.* at 472.

[129] Angela Hawken & Mark Kleiman, Managing Drug Involved Probationers with Swift and Certain Sanctions: Evaluating Hawaii's Hope 6-7 (Nat'l Inst. of Just., Document No. 229023, 2009), https://www.ojp.gov/pdffiles1/nij/grants/229023.pdf [https://perma.cc/9XDF-WPE9] (describing HOPE program).

[130] *Id.* (arguing that mild but consistent sanctions are more effective and less cruel than alternatives).

[131] *Id.* at 17-26 (comparing HOPE probationers' drug use to other probationers'); Angela Hawken et al., HOPE II: A Follow-up to Hawai'i's HOPE Evaluation 11-13 (Nat'l Inst. of Just., Document No. 249912, 2016), https://www.ojp.gov/pdffiles1/nij/ grants/249912.pdf [https://perma.cc/7EV4-CAM5] (following up on data after HOPE expansion).

[132] *See, e.g.*, TODD R. CLEAR & NATASHA A. FROST, THE PUNISHMENT IMPERATIVE 122 (2013) (calling HOPE "hottest new program in the field").

effect eventually.[133] By altering criminal justice sanctions to better correspond with behavioral incentives, proponents hoped to be able to reduce drug abuse without using big-stick carceral sentences.[134] The Project HOPE evaluation was backed by a theory of human nature that made its results seem broadly generalizable.[135] Observers described its rise as "meteoric": within a few years, "swift, certain, and fair" sanctioning had been adopted in at least 160 instances.[136] Again, the National Institute of Justice funded RCTs to try and replicate Project HOPE's success across five sites.[137] The results were not promising: "swift, certain, and fair" sanctioning did not offer any detectable

[133] *See* Philip J. Cook, Commentary, *Behavioral Science Critique of HOPE*, 15 CRIMINOLOGY & PUB. POL'Y 1155, 1157 (2016) ("Outside the realm of immediacy, there remains a tendency to discount the value of consequences according to just how far in the future they are expected.").

[134] Hawken & Kleiman, *supra* note 129, at 6 ("HOPE might represent a transformation in probation supervision: drastic reductions in rates of noncompliance achieved primarily through regular random drug testing combined with credible threats of low-intensity sanctions rather than revocations.").

[135] *Id.* at 6-7 (describing theory that criminals generally share certain personality traits and respond to certain incentives in predictable ways).

[136] J. C. Oleson, Commentary, *HOPE Springs Eternal: New Evaluations of Correctional Deterrence*, 15 CRIMINOLOGY & PUB. POL'Y 1163, 1167 (2016) ("HOPE has been replicated in 160 locations across 21 states . . . . Certainly, the rise of HOPE probation has been meteoric . . . ." (citations omitted)).

[137] *See* Pamela K. Lattimore et al., *Outcome Findings from the HOPE Demonstration Field Experiment: Is Swift, Certain, and Fair an Effective Supervision Strategy?*, 15 CRIMINOLOGY & PUB. POL'Y 1103, 1105 (2016) (discussing studies performed in Arkansas, Massachusetts, Oregon, and Texas); Daniel J. O'Connell, John J. Brent & Christy A. Visher, *Decide Your Time: A Randomized Trial of Drug Testing and Graduated Sanctions Program for Probationers*, 15 CRIMINOLOGY & PUB. POL'Y 1073, 1075 (2016) (discussing study performed in Delaware).

improvements over the status quo.[138] While jurisdictions may continue to operate in a HOPE-like fashion, the balloon of optimism has largely deflated.[139]

Few replication attempts have been as exhaustive as the HOPE ones, in part because few initial studies have been as compelling. In 2017, an article showed that a program teaching youths to "think slow" instead of impulsively responding led to a substantial decline in violent arrests and gains in high school graduation rates.[140] This well-executed study received a lot of attention and helped inspire the creation of Barack Obama's "My Brother's Keeper" ("MBK") initiative.[141] However, a follow-up article shows that this effect was mostly

[138] *See* Frances T. Cullen, Travis C. Pratt & Jillian J. Turanovic, Commentary, *It's Hopeless: Beyond Zero-Tolerance Supervision*, 15 CRIMINOLOGY & PUB. POL'Y 1215, 1223 (2016) ("[T]he future of HOPE ultimately lies in the data. At present, the evaluation research shows that this intervention has weak-to-null effects."). The original success and subsequent failure to replicate has sometimes been credited to the motivation and enthusiasm of the original judge behind Project HOPE. Janet Davidson, George King, Jens Ludwig & Steven Raphael, Managing Pretrial Misconduct: An Experimental Evaluation of HOPE Pretrial 69-70 (Jan. 2019) (unpublished manuscript) (available at https://gspp.berkeley.edu/assets/uploads/research/pdf/HOPE_final_evaluation_January_2019.pdf [https://perma.cc/APA8-TVHP]) (acknowledging different results may arise from original judge's skill in working with program individuals). A more recent RCT was conducted in Oahu with the original judge. *Id.* at 3 (describing RCT conducted between September 2014 and August 2016). The authors found some tentative successes: the treatment group had fewer failed drug tests, and fewer arrests involving a new criminal charge. *See id.* at 4, 6 (showing 21-30% lower drug test failure rate for treatment group members, and 41% lower rate of arrests involving new criminal charges). However, the overall arrest rate and the number of jail days served was equivalent across treatment and control. *Id.* at 5-6. The outstanding successes of the original study were not replicated, even with the same place and same group of people. The study remains unpublished.

[139] At least, it has deflated among many researchers. *See, e.g.*, PAMELA K. LATTIMORE, DEBBIE DAWES, DORIS L. MACKENZIE & GARY ZAJAC, NAT'L INST. OF JUST., EVALUATION OF THE HONEST OPPORTUNITY PROBATION WITH ENFORCEMENT DEMONSTRATION FIELD EXPERIMENT (HOPE DFE), FINAL REPORT 228 (2018) ("HOPE probation has been widely promoted and adapted as a means for substantially improving probation outcomes while generating cost savings. The findings of this rigorous four-site randomized controlled trial suggest otherwise."); Cullen et al., *supra* note 138, at 1222 (recommending end to promotion of Project HOPE as empirically supported model). However, jurisdictions continue to adopt this practice and some advocates continue to support it. *See* Francis T. Cullen, Travis C. Pratt, Jillian J. Turanovic & Leah Butler, *When Bad News Arrives: Project HOPE in a Post-Factual World*, 34 J. CONTEMP. CRIM. JUST. 13, 25-28 (2018) (describing advocates' arguments for promoting Project HOPE despite null findings from follow-up RCTs).

[140] Sara B. Heller et al., *Thinking, Fast and Slow? Some Field Experiments To Reduce Crime and Dropout in Chicago*, 132 Q.J. ECON. 1, 4 (2017) [hereinafter Heller et al., *Thinking, Fast and Slow?*] (showing 45-50% reduction in violent-crime arrests and 12-19% increase in high school graduation rates).

[141] *See My Brother's Keeper: Seven Years of Walking Alongside Youth and Communities*, MBK ALLIANCE (Feb. 26, 2021), https://www.obama.org/mbka/mbk-stories/mbk-7-years-

prevalent in the earliest cohort analyzed: effects for subsequent cohorts were close to zero and statistically insignificant.[142] The authors attribute this decline in effect to the possibility that, as the program scaled up, the skill and quality of the counselors declined.[143]

Note that this disappointing follow-up result would be hard to discover. While the original success was published in economics' most prestigious journal and received widespread media attention, the subsequent failure to replicate is mentioned only tangentially in the back pages of an unpublished working paper on a different topic.[144]

One of the few interventions in which success has been replicated is city-sponsored summer job programs for teens. Numerous RCTs have found that summer employment reduces criminal justice involvement.[145] Some of the studies suggest that effects persist past the summer of employment, implying that the effect is not entirely caused by keeping kids busy.[146] This is an important

youth-and-communities/ [https://perma.cc/NW5E-K22R] (attributing MBK's mentorship emphasis to "fact that research shows mentors can have tremendous impact on absenteeism, social-emotional growth, [and] school performance"); Nissa Rhee, *In Chicago, Can 'Thinking Slow' Prevent Crime?*, CHRISTIAN SCI. MONITOR (Nov. 5, 2016), https://www.csmonitor.com/EqualEd/2016/1105/In-Chicago-can-thinking-slow-prevent-crime [https://perma.cc/2NGL-YHRR] (reporting MBK was partially inspired by BAM, subject of 2017 study).

[142] *See* Bhatt et al., *supra* note 95, at 9 fig.2 (showing impact on school engagement and arrests approaching zero in 2013-14 and 2015-16 studies).

[143] *Id.* at 10.

[144] The original study was published in the *Quarterly Journal of Economics*. Heller et al., *Thinking, Fast and Slow?*, *supra* note 140, at 4. The subsequent failure to replicate was only mentioned in an unpublished working paper advocating for focusing on interventions with large scope. *See* Bhatt et al., *supra* note 95, at 8-10 (describing failure to replicate original study).

[145] *See* Alicia Sasser Modestino, *How Do Summer Youth Employment Programs Improve Criminal Justice Outcomes, and for Whom?*, 38 J. POL'Y ANALYSIS & MGMT. 600, 602 (2019) (finding that Boston Summer Youth Employment Program reduced violent crime by 35%); Alexander Gelber, Adam Isen & Judd B. Kessler, *The Effects of Youth Employment: Evidence from New York City Lotteries*, 131 Q.J. ECON. 423, 426 (2016) (finding that New York City Summer Youth Employment Program participation decreased probability of incarceration); Heller, *Summer Jobs Reduce Violence*, *supra* note 99, at 1219 (finding that summer jobs program for Chicago youth decreased violence by 43% over sixteen months); Sara Heller, *When Scale and Replication Work: Learning from Summer Youth Employment Experiments* 17 (Nat'l Bureau of Econ. Rsch., Working Paper No. 28705, 2021), https://www.nber.org/system/files/working_papers/w28705/w28705.pdf [https://perma.cc/RYU8-4CNL] [hereinafter Heller, *When Scale and Replication Work*] (finding that participation in summer youth employment programs in Chicago and Philadelphia consistently reduced criminal justice involvement).

[146] Modestino and Heller (2014) only looked at contemporaneous effects. *See* Modestino, *supra* note 145, at 600 (examining program effects over seventeen months); Heller, *Summer*

finding: summer jobs programs are relatively easy to implement and scale, and reductions in crime and criminal justice involvement are meaningful benefits. But note that there is no evidence that this intervention leads to wholesale change in youth trajectories. Summer jobs do not appear to increase average wages or employment after completion of the program, nor do they increase educational outcomes.[147] While some studies find impacts on arrests for violent crime, others find impacts only on drug or other minor offenses.[148]

Other jobs programs have achieved much less success. A large-scale RCT evaluated four different transitional jobs programs for recently-released prisoners across four different sites.[149] The programs did not increase regular (unsubsidized) employment during or after the program period, nor did they significantly affect key measures of recidivism over the two-year follow-up period.[150] An eight-site RCT of various programs designed to increase employment for "hard to employ" individuals, including those who have been justice-involved, found lasting employment effects for only one of the eight programs.[151] And a nationwide randomized evaluation of Job Corps, a program to boost employment for economically disadvantaged youth, showed that

Jobs Reduce Violence, *supra* note 99, at 1219 (examining program effects over eighteen months). Davis and Heller (2020) showed no effect in years two and three across two different cohorts. Jonathan M. V. Davis & Sara B. Heller, *Rethinking the Benefits of Youth Employment Programs: The Heterogeneous Effects of Summer Jobs*, 102 REV. ECON. & STAT. 664, 670 tbl.2 (2020). Gelber et al., however, showed impacts on adult incarceration; because the sample was almost entirely underage, this implies lasting crime effects. *See* Gelber et al., *supra* note 145, at 449 (showing 10% reduction in adult incarceration rate for minor participants). However, Gelber et al.'s results are only marginally statistically significant and might not stand up to a multiple hypothesis adjustment. *See id.*

[147] *See* Gelber et al., *supra* note 145, at 426 ("We do not find that youth employment has a positive effect on subsequent earnings or on college enrollment."); Davis & Heller, *supra* note 146, at 676 (noting that drop in violent crimes "occurs despite no detectable improvements in schooling, UI-covered employment, or other types of crime during the follow-up period"); *see also* Heller, *When Scale and Replication Work*, *supra* note 145, at 3 ("Neither city's participants show improvements in school engagement."). Davis and Heller did find some subgroup effects. Davis & Heller, *supra* note 146, at 676 (describing younger, academically engaged subgroup whose formal sector employment improved by 15%).

[148] *See* Modestino, *supra* note 145, at 602 (finding Boston Summer Youth Employment Program reduced violent crime arraignments by 35%); Heller, *Summer Jobs Reduce Violence*, *supra* note 99, at 1220 fig.1 (showing decrease in violent crime, but not other crimes); Davis & Heller, *supra* note 146, at 676; Heller, *When Scale and Replication Work*, *supra* note 145, at 3 ("In both cities, these changes were driven by significant decreases in arrests for drug and other non-violent, non-property crime arrests . . . .").

[149] ERIN JACOBS, RETURNING TO WORK AFTER PRISON: FINAL RESULTS FROM THE TRANSITIONAL JOBS REENTRY DEMONSTRATION 7 (2012) (describing study of transitional jobs programs in Chicago, Detroit, Milwaukee, and St. Paul).

[150] *Id.* at 9, 27.

[151] BUTLER ET AL., *supra* note 82, at 63 ("[O]nly PRIDE had impacts on regular employment that persisted over the full follow-up period.").

although earnings increased for several post-program years, gains ultimately were not sustained.[152]

Note that the interventions which get evaluated across multiple sites are those that, based on theory or prior research, are believed to be particularly promising. I focus on these not only because they have multiple evaluations, but also because one would expect these interventions to be the *most* likely to be effective. The idea that employment reduces crime is a staple theory in the social sciences; the proliferation of job-training programs is motivated by this core theory.[153]

Other widely-studied methods in the criminal justice space include hot spots policing, body-worn cameras, intensive probation, drug courts, and psychological interventions in prison. A number of RCTs have evaluated the impact of hot spots policing on the local areas that were randomly assigned to receive more attention from officers.[154] These studies show that the increased police presence leads to a small but statistically significant decrease in reported crime in the areas with increased policing.[155] The research on spillover effects of hot spots policing, or the long-term effects of hot spots policing on community well-being, is much less consistent.[156]

---

[152] Peter Z. Schochet, John Burghardt & Sheena McConnell, *Does Job Corps Work? Impact Findings from the National Job Corps Study*, 98 AM. ECON. REV. 1864, 1883 (2008) (finding that few, if any, program benefits accrued four years after program participation ended). The study also showed a reduction in arrests, although this is self-reported and therefore subject to reporting bias. *Id.* at 1874 (showing 29% arrest rate for treatment group, compared to 33% arrest rate for control group).

[153] *See* Christopher Uggen & Sarah K. S. Shannon, *Productive Addicts and Harm Reduction: How Work Reduces Crime—But Not Drug Use*, 61 SOC. PROBS. 105, 107 (2014) (providing overview of sociological and economic theories which suggest employment reduces crime).

[154] Anthony A. Braga, Brandon S. Turchan, Andrew V. Papchristos & David M. Hureau, *Hot Spots Policing and Crime Reduction: An Update of an Ongoing Systematic Review and Meta-Analysis*, 15 J. EXPERIMENTAL CRIMINOLOGY 289, 302 fig.4 (2019) (identifying thirty-five RCTs studying hot spots).

[155] *Id.* at 301 ("In this review, the quasi-experimental designs were associated with a larger within-group effect size (.171, $p<.001$) relative to the randomized controlled trial designs (.109, $p<.001$)."); *id.* at 298 ("[T]he overall effect size for these studies is .132 ($p<.001$); this would be considered a small mean effect size." (citation omitted)); *see also* Pamela Buckley et al., Does Hot Spots Policing Reduce Crime? An Alternative Interpretation Based on a Meta Analysis of Randomized Experiments 24 (unpublished manuscript) (on file with author) ("(1) On average, hot spots policing reduces crime; (2) The size of the average effect is between -0.046 and -0.051 standard deviation units . . . [T]he average effect is small and more studies did not find effects than did.").

[156] Christopher S. Koper, Cynthia Lum, Xiaoyun Wu & Tim Hegarty, *The Long-Term and System-Level Impacts of Institutionalizing Hot Spot Policing in a Small City*, 15 POLICING 1110, 1111 (2021) ("[I]t is not yet clear whether implementing a more widespread, systematic, and sustained preventative emphasis on hot spots in everyday police operations can produce

The impact of body-worn cameras in policing has also been widely studied with RCTs.[157] The studies show no clear improvement in either officer use of force, or assault or resistance against police officers.[158] Likewise, there have been many large and well-executed RCTs evaluating intensive probation, and there is no evidence that it reduces criminal activity relative to less intrusive supervision.[159] Recidivism effects for drug courts are also small and statistically insignificant when evaluated via RCT, as are psychological interventions in prisons.[160] In other words, the most widely studied strategies in criminal justice seem to have, at most, small and contested effects.[161]

Finally, the one area of causal inference research in which lasting or replicable effects are found somewhat more frequently is interventions made to the lives

large-scale aggregate reductions in crime for extended periods . . . . This is arguably one of the greatest remaining challenges to the study and practice of [hot spots policing]." (citations omitted)).

[157] Cynthia Lum et al., *Body-Worn Cameras' Effects on Police Officers and Citizen Behavior: A Systematic Review*, 16 CAMPBELL SYSTEMATIC REVS., no. 3, 2020, at 1, 2 (identifying at least thirty studies on effects of body-worn cameras).

[158] *Id.* at 27 tbl.8 (showing null effects found by RCTs). Body-worn cameras do have an impact on citizen complaints. *Id.* at 34. This may be due to a change in officer behavior, or it may be due to a reduction in frivolous complaints. *Id.* (suggesting reduction in complaints likely due to reduction in what officers feel are frivolous complaints, rather than significant changes in officer behavior).

[159] *See* Jennifer L. Doleac, *Study After Study Shows Ex-Prisoners Would Be Better Off Without Intense Supervision*, BROOKINGS (July 2, 2018), https://www.brookings.edu/blog/up-front/2018/07/02/study-after-study-shows-ex-prisoners-would-be-better-off-without-intense-supervision/ [https://perma.cc/K5CE-W6DC] (discussing four RCTs which found intensive supervision ineffective at reducing recidivism).

[160] *See* Ojmarrh Mitchell, David B. Wilson, Amy Eggers & Doris L. MacKenzie, *Assessing the Effectiveness of Drug Courts on Recidivism: A Meta-Analytic Review of Traditional and Non-Traditional Drug Courts*, 40 J. CRIM. JUST. 60, 66 tbl.4 (2012) (showing general recidivism rates in drug courts "considerably smaller in evaluations with higher levels of methodological rigor"); Gabrielle Beaudry, Rongqin Yu, Amanda E. Perry & Seena Fazel, *Effectiveness of Psychological Interventions in Prison to Reduce Recidivism: A Systematic Review and Meta-analysis of Randomized Controlled Trials*, 8 LANCET PSYCHIATRY 759, 768 (2021) (finding large-scale RCTs revealed no significant effect of psychological interventions on recidivism).

[161] A couple of recent RCTs have shown some success with place-based interventions (remediating vacant land or installing improved lighting), though it is yet to be seen whether these successes will replicate. Charles C. Branas et al., *Citywide Cluster Randomized Trial to Restore Blighted Vacant Land and Its Effects on Violence, Crime, and Fear*, 115 PNAS 2946, 2949 (2018) (finding interventions to restore blighted land "significantly reduced gun violence and other police-reported problems, such as burglaries and nuisances"); Aaron Chalfin, Benjamin Hansen, Jason Lerner & Lucie Parker, *Reducing Crime Through Environmental Design: Evidence from a Randomized Experiment of Street Lighting in New York City*, 38 J. QUANTITATIVE CRIMINOLOGY 127, 151 (2022) (finding "discrete environmental change brought about through tactical investment in enhanced street lighting can reduce violent and otherwise serious crimes appreciably in disadvantaged urban areas").

of youths.[162] Of all the employment-based interventions described above, the only effective ones (city-sponsored summer jobs) were targeted at young people.[163] A multisite RCT found that moving from high-poverty to low-poverty neighborhoods led to improvements in earnings and college attendance, but only for those who moved before the age of thirteen.[164] It did not, however, lead to any reduction in arrest rates.[165] Several early RCTs found that providing subsidized preschool for low-income families led to long-term increases in economic and health outcomes and decreases in crime.[166] However, scaling up

[162] *See, e.g.*, Lauren Bauer, *Does Head Start Work? The Debate Over the Head Start Impact Study, Explained*, BROOKINGS INST. (June 14, 2019), https://www.brookings.edu/blog/brown-center-chalkboard/2019/06/14/does-head-start-work-the-debate-over-the-head-start-impact-study-explained/ [https://perma.cc/8FT5-9P4B] (discussing impact of federal early childhood education program); Eric Chyn & Lawrence F. Katz, *Neighborhoods Matter: Assessing the Evidence for Place Effects* 7 (Nat'l Bureau of Econ. Rsch., Working Paper No. 28953, 2021), https://www.nber.org/papers/w28953 [https://perma.cc/8FT5-9P4B] (discussing impact of moving to wealthier neighborhoods in early childhood); Thomas Feucht & Tammy Holt, *Does Cognitive Behavioral Therapy Work in Criminal Justice? A New Analysis from Crimesolution.gov*, NAT'L INST. JUST. J. (May 25, 2016), https://www.ojp.gov/pdffiles1/nij/249825.pdf [https://perma.cc/FZ5M-P9K2] (discussing impacts of cognitive behavioral therapeutic intervention programs on criminal justice system).

[163] *See supra* notes 145-48 and accompanying text.

[164] Chetty et al., *supra* note 96, at 857-58 ("Children whose families were assigned to the Section 8 voucher group before they turned 13 generally have mean outcomes between the control and experimental group means. . . . The MTO treatments had very different effects on older children—those between 13-18 at RA . . . . The point estimates suggest that, if anything, moving to a lower-poverty neighborhood had slightly *negative* effects on older children's outcomes.").

[165] LISA SANBONMATSU ET AL., NAT'L BUREAU ECON. RSCH., MOVING TO OPPORTUNITY FOR FAIR HOUSING DEMONSTRATION PROGRAM: FINAL IMPACTS EVALUATION 257 (2011) ("We find no statistically significant effects of [Moving to Opportunity Program] on violent crime arrests in the long term data."). However, the researchers also concluded, "The one outcome for which we do see at least some hints of more pronounced impacts in the long-term data than in the interim data is with declining arrest rates for drug distribution among the [Moving to Opportunity] treatment groups compared to controls." *Id.* at 258.

[166] *See* Frances A. Campbell, Craig T. Ramey, Elizabeth Pungello, Joseph Sparling & Shari Miller-Johnson, *Early Childhood Education: Young Adult Outcomes from the Abecedarian Project*, 6 APPLIED DEVELOPMENTAL SCI. 42, 52 (2002) (finding that "[i]ndividuals assigned to the preschool treatment group had, on average, significantly higher cognitive test scores as young adults, . . . they earned higher scores on tests of reading and mathematical skills, they attained more years of education, they were more likely to attend a 4-year college or university, and they were less likely to become teen parents"); LAWRENCE J. SCHWEINHART, THE HIGH/SCOPE PERRY PRESCHOOL STUDY THROUGH AGE 40: SUMMARY, CONCLUSIONS, AND FREQUENTLY ASKED QUESTIONS 2 fig.1 (2005) (showing long-term benefits to economic and educational attainment and decreases in arrests). Subsequent re-evaluations of the data found beneficial effects only for girls. Michael L. Anderson, *Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry*

these early successes proved difficult: evaluations of the Head Start program have been mixed and equivocal.[167] Despite some occasional successes, we are still far from a thorough understanding of when childhood interventions will be successful, let alone how to make them scale.

III. THE STRUCTURE OF THE SOCIAL WORLD

We now have a battery of evidence. The evidence is imperfect; it is filtered through the human process of scientific research. But, as outlined in Section I.C., we understand this filtration process and the biases it produces. We know it tends to inflate results, making an intervention look more successful than it was.[168] Thus it is particularly striking that published studies—those that have made it through the filter—rarely find that the intervention was successful.

I argue in Section A that we learn something important from this. The hundreds of RCTs conducted in criminal justice should inform our beliefs about the structure of the social world. The evidence pertains most directly to questions answered and answerable by RCTs in the criminal legal space. But, as I argue in Section B, that doesn't mean the evidence is uninformative outside of that space.

A. *Stabilizers, Cascades, and Complexity*

A wide-lens view on more than fifty years of RCTs in the criminal legal space reveals a few common themes: most interventions don't work, and the ones that do tend not to replicate well in other settings. While this may be disappointing from the perspective of learning how to engineer social change, it teaches us something valuable about the structure of the social world. Namely, it teaches us that the social world is full of what I call "stabilizers" and short on what I refer to as "cascades." Note that I am not trying to *explain* my empirical claim by introducing these concepts. Rather, I am simply restating it in more abstract terms. Language has power, and, for me, this abstraction has helped me see the world in new ways.

Stabilizers are the set of socioeconomic forces that resist externally-imposed change. Imagine an orange in a large bowl. One can push the orange up the side of the bowl, but as soon as you let go, it rolls back to the bottom. A job-training program may provide a temporary job as well as some job-relevant skills. But whatever socioeconomic forces made it hard for that individual to find a job in

Preschool, and Early Training Projects, 103 J. AM. STAT. ASS'N 1481, 1482 (2008) ("The results demonstrate that early interventions . . . significantly improve later-life outcomes for females, . . . but that treatment effects are modest or nonexistent for males . . . .").

[167] This may have been because many youths in the control group simply attended a different preschool program. Patrick Kline & Christopher R. Walters, *Evaluating Public Programs with Close Substitutes: The Case of Head Start*, 131 Q.J. ECON. 1795, 1796 (2016) (suggesting conclusion that "Head Start is ineffective" may have been premature in part because "roughly one third of the [Head Start Impact Study] control group participated in alternative forms of preschool").

[168] *See infra* Section I.C.

the first place—for example, a society in which access to opportunity is deeply segregated—prove to be powerful inhibitors. After the program is over, the participant returns to the place they would have been absent the intervention.

The orange analogy has some strengths, but it implies a return to stasis and a direct force-counterforce dynamic. Stabilizing forces don't necessarily need to embody either. Consider another analogy: the tides. When the tide is pulling out to sea, the flotsam and jetsam will be carried along with it. A gentle cross breeze might create some eddies but will have little impact on the overall direction of flow. If change is governed by widescale social forces, then the interventions evaluated via RCT might be like this gentle breeze: ultimately irrelevant.

Stabilizers are closely related to a type of social change that I refer to as "cascades." Cascades are forces that magnify small changes, that turn a small intervention into a large and lasting effect. Consider, again, the example of the job-training program. Theoretically, a job-training program could launch a cascade. It could help recently released prisoners secure employment. Employment could then help secure housing, which could then create independence and security, which could in turn prevent drug use and other unhealthy behaviors, and so on and so forth until the person is reintegrated as a thriving member of society.

A cascade is defined by the idea that a small but well-timed intervention leads to a cycle of change that accumulates over time and affects many areas of one's life. Cascade narratives can be very compelling. But that doesn't mean they are true. RCTs teach us that very few interventions launch such a virtuous cycle of accumulating benefits. The social scientists' holy grail—the small, inexpensive intervention with large, widespread, and lasting gains—appears to be mostly myth.

Occasionally, however, someone claims to have found such a holy grail. Some intervention is demonstrated to be highly successful in one setting: so much so that other jurisdictions try to mimic their success. Unfortunately, the process identified in one setting rarely ports well to others.[169] Such instances suggest that causal processes are highly reliant on specific contextual conditions.[170] As a program expands, it may have trouble recruiting and training staff to the same skill level.[171] The original success could be due to the charisma

[169] *See supra* notes 112-43 and accompanying text.

[170] Of course, this could also mean that the original result was a false positive, meaning an instance in which luck (or researcher manipulation) made an unimpactful intervention look more successful than it was.

[171] *See, e.g.*, Bhatt et al., *supra* note 95, at 10 (explaining potential difficulty of recruiting more skilled counselors as Chicago's Becoming a Man ("BAM") program expanded); Sampson, *supra* note 89, at 494 ("It is not the location or population differences so much as that once a policy takes effect the rules of the game change, possibly inducing system-level changes.").

of a key figure, without whom the intervention has much less impact.[172] Or it could be contingent on a particular set of background conditions—low crime rates, low unemployment rates, etc.[173] If success relies on a particular alignment of the stars, a causal process detected in one setting teaches us little about what will happen once the stars shift.

B. *Scope of the Claim*

The evidence presented in Part II is derived from a particular set of studies conducted in a particular time and place. My claim is that these studies teach us something broader about the structure of the social world. This is an inductive argument. Like any inductive argument, the extent to which one can extrapolate beyond a particular time and place is up for debate. Inductive arguments find their surest footing when applied to settings similar to where the data was derived. But that doesn't mean they are uninformative when applied further afield. I discuss generalizability along several dimensions in this Section.

First, I want to reiterate that this is a claim about the nature of the social world and does not extend to physics or biology. Medical research, for instance, has clearly shown that limited scope interventions (e.g., drugs or vaccines) can have large and widely replicable effects. Fields such as public health, which straddle medical and social sciences, may be exempt for similar reasons.

1. Does the Claim Apply Outside of the Criminal Legal Space?

Perhaps there is something unique about the *people* studied by RCTs in the criminal legal space that could limit generalizability. For instance, the people in these studies often come from marginalized groups; they have little access to society's wealth and resources.[174] Do these factors make their life trajectories particularly difficult to change? Would interventions made to the lives of better-resourced individuals make more of a difference?

---

[172] That's one hypothesis for the failure to replicate Project HOPE. Stephanie A. Duriez et al., *Is Project HOPE Creating a False Sense of Hope? A Case Study in Correctional Popularity*, FED. PROB., Sept. 2014, at 57, 67-68 (2014), https://www.uscourts.gov/sites/default/files/78_2_7_0.pdf [https://perma.cc/YZV8-XEFP] (noting policymakers "might have paused to wonder whether a program based on a limited theory of crime that has rarely succeeded in producing effective interventions . . . might have only limited effects and not be effective in courtrooms not led by a charismatic judge").

[173] Cartwright and Deaton refer to this as the "transportation" problem. Angus Deaton & Nancy Cartwright, *The Limitations of Randomised Controlled Trials*, VOXEU: CEPR (Nov. 9, 2016), https://voxeu.org/article/limitations-randomised-controlled-trials [https://perma.cc/2U5J-4ZA3] ("Causal effects depend on the settings in which they are derived, and often depend on factors that might be constant within the experimental setting but different elsewhere. Even the direction of causality can depend on the context.").

[174] *See, e.g.*, Alicia H. Munnell, *Lessons from the Income Maintenance Experiments: An Overview*, *in* LESSONS FROM THE INCOME MAINTENANCE EXPERIMENTS 1-3 (Alicia H. Munnell ed., 1986) (outlining income experiments performed on welfare recipients).

At first glance, those who occupy more privileged places in society may appear to have a greater capacity to respond to interventions. They have greater access to resources to help them advance; their higher levels of education may give them enhanced knowledge about opportunities and consequences. But to the extent that more advantaged groups have a greater capacity to fulfill their wishes, *they can accomplish this on their own*. The equilibrium state of advantaged groups looks different from the equilibrium state of less advantaged groups. But, like everyone else, they are living the best life they can create for themselves given the opportunities provided to them, the knowledge they have, and the set of skills and burdens they carry. If it was easy for them to have made a meaningful improvement, they would have done so already.

The relevant question here is the extent to which externally imposed and short-term changes to circumstances—interventions—lead to meaningful, lasting, and replicable changes in people's life trajectories. I don't see why this would be more likely with an advantaged population than a disadvantaged one. To the extent that advantaged groups have greater capacity, they are starting from an equilibrium in which they are closer to having achieved their goals. To the extent that their goals remain out of reach, that's due to the constraints relevant to them. And there is no clear reason to assume that these constraints are more easily cleared away than those relevant to less advantaged populations.

One could also speculate that the claim applies only to the set of *interventions* relevant to criminal justice reform. Again, I don't think this is the case. For practical purposes, I've limited my empirical analysis to one substantive area, but I believe that, at least in broad strokes, the claim extends beyond this setting.

My beliefs are informed by my knowledge of the empirical literature in other fields. I was trained by development economists and I frequently attend conferences and collaborate with economists studying education, health, and labor. In these fields also, interventions evaluated via RCT rarely find large or lasting benefits. Microcredit (loaning small amounts of money, often to women) was, for many years, the darling of the development world. Eventually, it was shown to have little to no net benefit in most places.[175] Health insurance, randomly allocated via lottery, was shown to increase healthcare usage and reduce bills sent to a collections agency.[176] Yet, it had no statistically significant effect on physical health or labor market outcomes.[177] The fact that most interventions in the social world have little impact is so ubiquitous that it has

[175] *See* Rachael Meager, *Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments*, 11 AM. ECON. J.: APPLIED ECON. 57, 86 (2019) (arguing empirical evidence across studies indicates impact of micro-credit interventions "d[id] not transform the lives of poor households in measurable ways, as was initially hoped").

[176] *See Oregon Health Insurance Experiment-Results*, NAT'L BUREAU OF ECON. RSCH., https://www.nber.org/programs-projects/projects-and-centers/oregon-health-insurance-experiment/oregon-health-insurance-experiment-results [https://perma.cc/62A3-S4MV] (last visited Nov. 10, 2023) (summarizing findings of healthcare expansion experiment).

[177] *See id.*

been dubbed the *Iron Law of Evaluation*: "The expected value of any net impact assessment of any large-scale social program is zero."[178]

To the extent that RCTs in other fields may find effects more frequently, this could be partly because they are more interested in quantifying a direct effect rather than establishing an indirect or longer-term effect. For instance, a group of large-scale RCTs conducted in the 1960s-1980s evaluated the impact of providing a guaranteed income to low-income individuals.[179] One of the primary goals of these studies was to quantify *how much* less people would work, with few questioning whether there would be a labor supply effect at all.[180] As expected, all studies found that cash transfers led to lower labor supply.[181] But when it came to more indirect outcomes—health, consumption habits, etc.—the researchers found that, overall, "the lives of recipients were not altered dramatically by the payments offered in the experiments."[182]

In sum, I expect my claim about the structure of the social world applies to many different types of people and many different types of interventions—at least where the proposed mechanism of causal influence is indirect or convoluted. The more speculative it sounds, the less likely it is there will be a robust and replicable causal relationship.

2. Does the Claim Apply Beyond the Set of Questions Answered and Answerable by RCTs?

My claim applies most directly to the questions answered by RCTs. As discussed in Section I.D, this imposes some important constraints to the scope of the claim. One is that RCTs tend to focus on questions that aren't *a priori* obvious. Although there is certainly a gray area in what constitutes "obvious," there are also many clear-cut examples. One does not need an RCT to evaluate whether providing food to the hungry fills bellies. Outcomes that are the direct, mechanical effect of a reform or intervention are generally too obvious to fall within the scope of my claim.

Another constraint is that the interventions evaluated by RCTs must be isolable and are generally of limited scope.[183] However, these constraints are not unique to RCTs. All causal inference methods embody these constraints to a certain degree.[184] This is because, at its heart, empirical causal inference entails

[178] *See* Rossi, *supra* note 2, at 4.

[179] Munnell, *supra* note 174, at 1 (discussing findings regarding effectiveness of U.S. public welfare programs from 1960s through 1970s).

[180] *Id.* at 1 (describing, as motivation for studies, "widespread fear that a guaranteed income would reduce the work effort of poor breadwinners"); Heckman, *Randomization and Social Policy*, *supra* note 89, at 4 ("The policy question was whether imposition of [guaranteed income payments] would substantially reduce labor supply.").

[181] *See* Burtless, *supra* note 76, at 35 (summarizing statistical findings).

[182] Munnell, *supra* note 174, at 8.

[183] *See supra* notes 74-78 and accompanying text.

[184] ANGRIST & PISCHKE, *supra* note 25, at xiii.

comparing two groups that differ in treatment status but are similar in all other relevant ways.[185] This helps ensure that any difference in outcomes can be attributed to the treatment as opposed to some other source. In order to have two groups who are otherwise similar except for some intervention, that intervention needs to be isolable from other factors. And isolability often means that the intervention is of limited scope.

Consider an example. Although incarceration is not something that is usually evaluated via RCT due to ethical considerations, its impacts can be evaluated using a natural experiment, or an instance in which the natural processes of the social world mimic certain aspects of an RCT. For instance, defendants may be randomly assigned to strict or lenient judges.[186] Or the sentence guidelines schema may create discontinuities in the sentencing recommendations: those scoring just above a cutoff receive harsher sentences than those right below, despite being similar in most other ways.[187] These types of natural experiments create otherwise similar defendants who vary in the length of their sentence.[188]

Idiosyncratic variation like this is usually somewhat limited. The treatment group might get, say, twelve-month-long sentences, while the control group only gets four months.[189] I don't mean to say that an additional eight months of incarceration isn't anything—to the incarcerated person, this is certainly a big deal. But it's limited relative to many of the research questions people want answered. Natural experiments generally don't allow us to, say, identify the impact of a five-year prison sentence compared to the path life would have taken absent any criminal justice involvement at all. There really is no good way to empirically identify the impact of a five-year prison sentence versus no criminal justice involvement at all. A researcher might attempt to use a method like matching to identify two groups who differ in their extent of justice involvement but are similar in the other characteristics captured in the data. But data is limited, and these two groups almost certainly differ in some important ways that are unobservable by the researcher. Causal inference for this type of

[185] *Id.* ("The [econometrics] craft uses data to get to *other things equal* in spite of the obstacles—called selection bias or omitted variables bias—found on the path running from raw numbers to reliable causal knowledge.").

[186] *See, e.g.*, Charles E. Loeffler & Daniel S. Nagin, *The Impact of Incarceration on Recidivism*, 5 ANN. REV. CRIMINOLOGY 133, 147-49 (2022) (summarizing findings of judge instrumental variable-based studies, intended to account for selection bias, on impact of imprisonment on recidivism rates).

[187] *Id.* (summarizing findings of sentence regression discontinuity studies, intended to account for selection bias, on impact of imprisonment on recidivism rates).

[188] Most studies find that variation in exposure to postconviction incarceration on the scale evaluated via natural experiment has little impact on recidivism after release. *Id.* at 147.

[189] *See* John Eric Humphries, Aurelie Ouss, Kamelia Stavreva, Megan T. Stevenson & Winnie van Dijk, Conviction, Incarceration, and Recidivism: Understanding the Revolving Door 36 (July 12, 2023) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4507597 (summarizing findings of recidivism study based on quasi-random judge assignment).

research question relies heavily on assumptions about how unobserved differences affect the outcomes.

The scope of my claim is thus not just limited to interventions evaluable via RCT, it's limited to interventions evaluable via rigorous method of empirical causal inference. Both impose a strict "otherwise similar" requirement that effectively means that the interventions evaluable are of limited scope. I would argue, moreover, that the human process of learning through experience also rests heavily on the "otherwise similar" constraint. When we form expectations about the impact of some intervention, our expectations have stronger empirical ground when experience allows us to compare groups or instances that are otherwise similar except for the intervention.

To sum up: one can imagine a sliding scale between research questions that are easily evaluated using empirical causal inference methods and those that are not. At the far end of this continuum lies the type of deep, systemic reform for which the "otherwise the same except for X" thought experiment really starts to falter. Prison abolition, for instance, lies on this far end of the scale. There are no natural experiments that one could leverage to identify the causal effect of prison abolition. Even if there were, it wouldn't be possible to disentangle the causal effect of prison abolition from the entire societal transformation that would need to occur before prison abolition was even a possibility.

The arguments made in this Article say nothing about reform on that scale. Rather, they apply most directly to questions in which the basic framework of empirical causal inference is relevant: the impact of isolable and limited scope interventions.

IV. IMPLICATIONS

The primary goal of this Article is to build and support the claim that, when it comes to the type of limited-scope interventions evaluated by RCTs, the social world is full of stabilizing forces that resist change. This claim has a variety of implications. Because I hope to keep this Article brief, a full discussion of them is beyond scope. However, I provide a brief sketch below. These are not fully developed arguments, but rather observations that I hope will lead to more research in the future.

I begin by noting that, at least within the confined scope discussed above, a pervasive view about the structure of the social world appears to be at least partially a myth. This myth forms a background assumption for many people in policy and academia, and it has been a dominant paradigm for reform over the last several decades. Setting it aside opens new doors for thinking about how to achieve social change. It also provides an opportunity to reflect on our methods of generating knowledge about the world, and the processes impeding it.

If the reader prefers to interpret these implications as applying solely to the criminal justice domain, I am content with that. While this is not the stance I take, I acknowledge that the handful of paragraphs I included to argue that the

scope is wider may be insufficient to convince a skeptical reader.[190] At the very least, I hope this conversation will provoke similar discussions in other domains.

A. *Myth*

There is a common view about the structure of the social world that I refer to as the *engineer's view*. Under the engineer's view, the causal structure of the social world can be mapped using RCTs and other scientific methods, and, once mapped, it can be manipulated to achieve social goals. Certain interventions yield such consistent and replicable success that they can be labeled "best practices." And meaningful reform can be achieved with reduced risk and uncertainty because the interventions have been rigorously evaluated before scaling up.

I use the phrase "engineer's view" as a term of art in this Article. In my usage, the engineer's view embodies both a *substantive* claim about the structure of the social world (i.e. that its structure is amenable to manipulation and control) and an *epistemological* claim about our ability to reliably predict the impacts of our reforms.

The engineer's view of the social world is widespread among academics, philanthropic organizations, and think tanks. Consider how Arnold Ventures, a philanthropic organization active in the criminal justice space, describes their mission:

> We focus on correcting system failures through evidence-based solutions. Viewing philanthropy as an engine of innovation, we identify problems, rigorously research them, and search for answers. Once an idea is tested, validated, and proven efficacious, we fund policy development and technical assistance to create change that outlasts our funding.[191]

Or consider how the National Institute of Justice, the research arm of the DOJ, describes its work:

> Science supports corrections agencies and the larger criminal justice system by delivering precise, reliable processes capable of generating consistent, repeatable outcomes.[192]

I don't mean to oversimplify. People's viewpoints are complex and can't be reduced to a single framework. Even if people sometimes seem to adopt the engineer's view, their full beliefs are likely to be much more nuanced and

---

[190] My claim does not apply to medical interventions, and it may not apply to quasi-medical domains like social health. Interventions on the physical body, such as drugs, diet, or hand washing, have a well-documented impact on many outcomes. Thus, social policies targeting health and the physical body, such as vaccine mandates or free malaria nets, may also have an impact.

[191] *Arnold Ventures: Strategy*, GLOB. JUST. RES. CTR., https://globaljusticerc.org/arnold-ventures/ [https://perma.cc/RY7B-ESQ7] (last visited Nov. 10, 2023). For full disclosure, Arnold Ventures has generously funded other research of mine.

[192] Nat'l Inst. of Just., *supra* note 9, at 12.

multifaceted. Nonetheless, the engineer's view forms a pervasive and influential theme in discourse. People frequently speak and behave as if the world works in an engineerable fashion. This can be found in the emphasis on using RCTs to identify "what works," as if this is some sort of universalist answer, a part of the inner functioning of the machine.[193] It can be seen in the calls to make policy "evidence-based,"[194] which is often (and often falsely) taken to mean "proven effective."[195] And it shows up in dialogues about engaging in research before "scaling up."[196]

The engineer's view may have particular traction among policy wonks and academics, but its roots go deeper than that. Almost everyone thinks in engineering terms, at least occasionally. People use narrative-based structures of cause and effect to interpret their own lives or the lives of people around them. At its heart, the engineer's view is just a story for how the world works and a proposal for how to change it. And such storytelling can be very compelling. Consider, for example, this hypothetical pitch:

> People released from prison are extraordinarily vulnerable. They often have no money, no home, and little prospect for employment with a felony conviction on their record. This is a pivotal time. If they can find a job, they can begin the process of reestablishing themselves within the community. If they can't, they are likely to be back in prison within a few months. This is an opportunity for us to intervene! We need to invest resources to help recently released individuals find jobs. We need to connect them with employers and support them in creating a resume or prepping for interviews. We should even subsidize their employment for a few months to help establish good work habits and add recent employment to their resume. A program like this has the potential to yield very high dividends, since it will launch people onto a new life trajectory with increased rates of employment and reduced rates of crime and incarceration.

You don't need to be a policy wonk to find such a pitch convincing. It creates a narrative to explain a social problem: people wind up back in prison because they have trouble getting jobs. Out of this narrative flows a seemingly obvious

---

[193] See, for example, the mission statement of the Chicago Crime Lab, a preeminent consortium of academics and researchers doing work in the criminal justice space: "Crime Lab staff partner with civic and community leaders to generate evidence on what works to tackle crime, violence, and the collateral costs [of] the criminal justice system." *Urban Labs: Crime Lab*, UCHICAGO URB. LABS, https://perma.cc/PA3K-L4KZ (last visited Nov. 10, 2023); *see also* JUST. TECH LAB, http://justicetechlab.org [https://perma.cc/2N9Q-JHHU] (last visited Nov. 10, 2023) ("Our team figures out what works, to make our policies better.").

[194] Megan Stevenson, *Assessing Risk Assessment in Action*, 103 MINN. L. REV. 303, 312-14 (2018) (providing overview of evidence-based criminal justice movement).

[195] See, for example, the tagline of the Justice Tech Lab: "Finding Effective, Scalable Solutions to Criminal Justice Problems." JUSTICE TECH LAB, http://justicetechlab.org [https://perma.cc/2N9Q-JHHU] (last visited Nov. 10, 2023).

[196] *See id.*

solution: help people find jobs and they won't wind up back in prison. This pitch is an example of the engineer's view in action. It presumes a mechanistic structure that can be predictably manipulated to achieve social goals.

As an empirical economist, the engineer's view is the water in which I've swum for the last ten years. That's not to say that people generally discuss things in such terms. Almost no one in my field talks about the structure of the social world in the broad, abstract manner that I have been employing here. Nor do most researchers adopt a tone quite as confident as some of the policy organizations I've quoted above. But the engineer's view can be found almost everywhere. It shows up in the reward structure of our field. The empirical research that gets celebrated is that which purports to successfully map some quadrant of the causal machine, that shows how an intervention successfully changes an important outcome.[197] The engineer's view can also be seen in what research gets swept under the rug, i.e., interventions that had little success or replication attempts that failed to pan out.[198] The presumption that the causal structure of the social world is something that can be mapped through research and manipulated to achieve social goals is part of the dominant paradigm of social science.

But over fifty years of RCTs in the criminal legal space call the engineer's view into question. At least when it comes to questions answered and answerable by RCTs, the engineer's view appears to be mostly myth. That doesn't mean that human actions never have an impact, but rather that the type of discrete, limited-scope interventions that are the primary domain of empirical causal inference research generally have limited or nonreplicable impact.

What about interventions of a much larger scope, reforms that address many aspects of the social world at once, such as technological or social revolutions? Can people engineer the change they want to see with reform on this scale? Perhaps. I make no claim in this area, as I don't think RCT evidence is directly relevant. However, I don't think any empirical method provides clear evidence about the impact of change on this scale. You can't pilot test large-scale or systemic reform. We don't have good tools to map the causal structure of the social world when it comes to social or technological revolutions. Predictions depend more on theory, ideology, analogy, and assumption. And these predictions are likely to be as disputed as the grounds on which they are based.

B. *Social Change*

Although the engineer's view has broad influence, it is most closely associated with a movement referred to as "evidence-based reform." The central

[197] As an example, consider the difference in reception between the original BAM study, which made the intervention look wildly successful, and the subsequent study, which didn't. The former was published in a top journal and given widespread media attention, and the latter was mentioned only tangentially in a paper on another topic, and remains virtually unknown. *See supra* note 144 and accompanying text.

[198] *See supra* Section I.C.

idea in the evidence-based movement is to make the criminal legal system more effective by adopting practices backed by evidence, usually understood to mean RCTs or other quantitative social science research.[199] Evidence-based reform has been a central part of reform discourse for a couple of decades now, with both major political parties endorsing it.[200] The requirement that reform be evidence-based is sometimes even enshrined in law.[201]

I am not opposed to evidence. This is an evidence-based Article, in that I build my entire argument around evidence derived from RCTs. Nor am I opposed to the evidence-based reform movement, which encompasses much more than a fondness for RCTs.[202] But I do question the vision of social change embodied by the evidence-based movement. This vision of social change derives in part

[199] *See* Collins, *supra* note 5, at 416-18 (describing sort of evidence prioritized); Robert J. Sampson, Christopher Winship & Carly Knight, *Translating Causal Claims: Principles and Strategies for Policy-Relevant Criminology*, 12 CRIMINOLOGY & PUB. POL'Y 587, 588-89 (2013) ("[E]vidence-based policy has largely become equated with evidence from randomized controlled trials (RCTs).").

[200] Kara Gotsch, Opinion, *Law and Order Agenda Should Take Note of Bipartisanship's Results*, ATLANTA J. CONST. (Nov. 29, 2016), https://www.ajc.com/news/opinion/law-and-order-agenda-should-take-note-bipartisanship-results/iu7CG22rf1qdmeBu95DFRI [https://perma.cc/2B9C-UXEU] ("Lawmakers in both parties now support evidence-based policies that promote public safety, value opportunities for redemption and are cost-effective.").

[201] *See, e.g.*, First Step Act of 2018, Pub. L. No. 115-391, 132 Stat. 5194 (codified as amended in scattered sections of 18 U.S.C., 21 U.S.C., and 34 U.S.C.) (requiring implementation of several evidence-based reform programs).

[202] I am, however, sympathetic to many critiques of the evidence-based movement, particularly those that emphasize the subjective and normative aspects of what is purportedly a neutral, scientific approach to policy. *See generally* Ngozi Okidegbe, *Discredited Data*, 107 CORNELL L. REV. 2007, 2007 (2022) (asserting purportedly neutral pretrial algorithms reproduce inequities partly because they are built with data exclusively from "carceral knowledge sources," such as pretrial services agencies); Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59, 59 (2017) (describing various normative judgements developers make when creating tools used to predict recidivism risk); Collins, *supra* note 5, at 403 (arguing that evidence-based movement is political, with agenda that strengthens rather than challenges existing system); Bernard Harcourt, *The Systems Fallacy: A Genealogy and Critique of Public Policy and Cost-Benefit Analysis*, 47 J. LEGAL STUD. 419, 419 (2018) (arguing systems-analytic decision-making techniques, such as cost-benefit analysis, entail normative judgements and shape political outcomes); Klingele, *supra* note 4, at 537 (illustrating how many evidence-based practices originally intended to make the criminal justice system more humane can be used to further empower the penal state); Benjamin Levin, *Criminal Justice Expertise*, 90 FORDHAM L. REV. 2777, 2784 (2022) (discussing how different kinds of expertise, including that based on lived experience, should be valued when creating a framework for political decision-making regarding criminal justice); Jocelyn Simonson, *Police Reform Through a Power Lens*, 130 YALE L.J. 778, 789 (2021) (expounding theory of power-shifting based on policy proposals from social movements calling for police reform, and offering broader perspective from which scholars can measure success in such reform).

from the methodologies it embraces.[203] Epidemiologist Sharon Schwartz and her co-authors argue that a focus on RCTs means a focus on a type of reform that is inherently conservative.[204] Not conservative as in Republican-leaning, but conservative as in favoring the conservation of an existing structure or system.[205] RCTs require holding all else constant except the treatment. This "limits the focus to interventions that leave systems intact and change some element that is manipulable without doing 'damage' to the system."[206] A focus on RCTs means prioritizing a certain set of policies, policymakers, and experts.[207]

RCTs and other causal inference methods may be associated with a conservative approach to reform, but they do not necessarily *support* a conservative approach to reform. Rather, they are a *test* of the conservative approach. And the test shows that system-conserving changes rarely have much lasting effect.

If the approach to social change espoused in the evidence-based reform movement is unlikely to have much effect, what next? People are not going to give up their desire to make a meaningful difference in the world just because it isn't as easily engineerable as some had hoped. If limited-scope changes to one component of a complex system rarely have much lasting effect, then this leaves us with only a few options: (1) focus on interventions with immediate, direct net benefits; (2) continue engaging in limited-scope interventions with an acceptance that they are unlikely to have more than a small impact; or (3) try to implement change that goes beyond what is evaluable with RCTs. I discuss each in turn.

(1) Give a man a fish: There is an old cliché that if you give a man a fish, he will eat for a day; if you teach him how to fish, he will eat for a lifetime. Such sentiments form the basis of many of the interventions discussed in this study. These interventions, designed to give people the resources to thrive on their own, rarely have large or lasting impact. The cliché is wrong, at least when it comes to the limited-scope, systems-conserving interventions. However, there remains a straightforward and obvious way to ameliorate harm: simply give people what they need. If they are hungry, give them food. If they need shelter, give them a home. If they need work, give them a job.

---

[203] *See* Sampson et al., *supra* note 199, at 589 ("In the case of the causes of crime, classic criminological subjects such as poverty or subcultural values are typically considered root causes. Yet the turn toward causality and policy has pushed much of criminology away from this kind of focus.").

[204] Sharon Schwartz, Seth J. Prins, Ulka B. Campbell & Nicolle M. Gatto, *Is the "Well-Defined Intervention Assumption" Politically Conservative?*, 166 SOCIAL SCI. & MED. 254, 255 (2016) (noting RCTs, by definition, leave all but one thing as is, while more radical change involves many moving parts).

[205] *Id.*

[206] *Id.* at 256.

[207] *See id.* (describing well-defined interventions); Collins, *supra* note 5, at 410 (describing privileging knowledge of "expert").

(2) Uncertain incrementalism: The evidence presented here shows that limited-scope interventions rarely have large or lasting effects on the outcomes measured. However, they may have small effects—effects too small to be detected using statistical methods. Some people may consider this sufficient. A series of small tweaks to the system could eventually accumulate into meaningful change.

However, such a claim is speculative. While the statistical methods used to evaluate limited-scope interventions generally don't allow us to reject small gains, they also don't allow us to reject small losses. There is an inherent uncertainty to this type of incrementalism. We may hope that a series of small steps will allow us to travel in the right direction, but we cannot know for sure. We could just as easily be traveling in the wrong direction.

It's also possible that limited-scope interventions have the type of impact that is unmeasurable or difficult to measure. Again, I cannot rule that out. But I would argue that large changes to unobserved traits would often result in observable changes as well. For instance, if an intervention had a large and lasting shift on well-being, we might expect it also to decrease future arrests or increase future employment. The failure to find such effects makes the inference of unobserved gains much more speculative.

(3) Systemic reform: For those who desire larger-scale change, and for whom incrementalism is too slow and uncertain, there really remains only one option: systemic reform. This Article shows that limited-scope, isolable interventions rarely lead to meaningful change. Those who desire meaningful change must therefore seek interventions outside the scope of what is evaluable via RCT. This includes changes that are so multipronged and entangled that it is impossible to hold all else constant. This also includes changes that are so large in scope that experimental evaluation is infeasible.

Systemic reform is not an "intervention" in the way I've been using that term here. It's not something you can do on your own; it requires changing the hearts and minds of large numbers of people, as well as changing the concrete structural factors of our lived experience. It's hard to know what systemic reform will bring, not only because we cannot test its impact empirically, but because it's very hard to imagine a world that is otherwise the same as ours, while also being deeply, structurally different. When it comes to systemic reform, we are flying half-blind.

C. *What Is the Structure of the Social World?*

The evidence generated by RCTs helps us to reject one particular view on the structure of the social world: the engineer's view. But that leaves open a vast terrain of possibilities. Some readers of this Article have suggested my claim

supports a Hayekian view of social processes.[208] Indeed, Hayek's famous quote looks at first glance like a pithy comment on my Article: "The curious task of economics is to demonstrate to men how little they really know about what they imagine they can design."[209] If the social world doesn't follow a simple, mechanistic, and engineerable set of processes, perhaps it is better described as a Hayekian system of spontaneous order.[210] Of course, Hayek's critique was primarily addressed at large-scale social engineering attempts, not the type of limited-scope, systems-conserving reforms that RCT evidence speaks to.

Other readers have interpreted my claim to be Marxist (or, at least, consistent with a Marxian critique). After all, why would one expect limited scope interventions to have an effect if they leave unchanged the mode of production of material life, which is ultimately responsible for "determin[ing] the general character of the social, political and spiritual processes of life"?[211]

So which is it, Hayek or Marx?[212] Or some other view on the structure of the social world? I don't know, and I don't think RCTs provide much in the way of an answer. My goal is just to show that one prominent viewpoint—the engineer's view—is not supported by the evidence. This may create a natural moment of inquiry about what alternative view should replace it. I celebrate that inquiry but leave such contemplation to the reader.

D. *How Should We Learn About How To Achieve Desired Change?*

In a previous section, I proposed a hypothetical sliding scale between limited-scope-but-answerable research on one end and larger-scope-but-harder-to-answer questions on the other.[213] If the type of intervention evaluated via RCT tends not to have large or lasting effect, perhaps we should focus more of our research attention on the other end of the scale. Large-scale change clearly occurs, but why? What causes these seismic shifts?

While these types of big-picture questions are incredibly important, I don't think empirical causal inference research is currently in a good place to answer them. The incentives are too distortionary. Researchers know that their paper will only be successful if they show that whatever cause they evaluate has a

---

[208] FRIEDRICH A. VON HAYEK, THE FATAL CONCEIT: THE ERRORS OF SOCIALISM 7 (W.W. Bartley III ed., Univ. of Chi. Press ed. 1989) (1988) (criticizing socialist desire for "deliberate arrangement of human interaction by central authority").

[209] *Id.* at 76.

[210] Friedrich A. von Hayek, *Kinds of Order in Society*, NEW INDIVIDUALIST REV., Winter 1964, at 3, 4-5 (theorizing more complex social order develops spontaneously under certain conditions, rather than as result of people deliberately arranging social elements).

[211] KARL MARX, A CONTRIBUTION TO THE CRITIQUE OF POLITICAL ECONOMY 11 (N.I. Stone trans., Charles H. Kerr Publ'g Co. 1904) (1859).

[212] Thanks to Andrew Hayashi, Paul Mahoney, and Thomas Frampton for helping me see the interesting connections to Hayek and Marx.

[213] *See supra* Section III.B. This inverse relationship has been noted and discussed before, but I am unaware of its origins.

statistically significant effect.[214] Accordingly, many engage in questionable research practices to be able to find such an effect, and, with the greater degrees of freedom that come with these less rigorous research designs, they have more latitude to do so.[215]

But it's not just the fraudulent practices that are the problem. There is also a selection problem in terms of which early-stage projects a researcher decides to bring to completion. If a researcher begins to investigate a question and the results are statistically insignificant, the project will almost certainly be abandoned.[216] There is no stigma attached to this, it is freely discussed, and it is almost universal. Generally speaking, the only time a researcher will proceed with "null" (statistically insignificant) results is when it's an RCT or the type of natural experiment that closely mimics an RCT. In other words, statistically insignificant results will generally only get written up and published when the research question is on the limited-scope-but-answerable side of the sliding scale.[217] This creates massive distortions in the universe of published research on the larger-scope-but-harder-to-answer side of the scale. I currently find this type of research very hard to learn from.

I hope that one day these distortionary incentives will no longer exist. Part of my goal for this Article is to prompt the type of reflection that would spur such a change. In the meantime, I don't mean to write off all causal inference research. RCTs and other quasi-experimental methods may provide valuable theoretical insight that can help inform policy. With much larger samples, we may be able to identify interventions that have small but meaningful impact—or we may be able to identify the subgroups of individuals for whom the intervention matters most. That being said, identifying small effects or effects that differ across subgroups requires exponentially larger sample sizes. This creates important practical limits to what is likely to be accomplished via RCT.

Of course, there are a variety of other modalities through which we may be able to learn how to effect social change: qualitative research, theoretical research, descriptive quantitative research, and so forth. I've focused my discussion on empirical causal inference because that's my area of expertise, but that is certainly not the only potential path to knowledge. All modalities have their challenges and limits, but I believe they all have the potential to contribute valuable insight on the mechanisms of social change.

---

[214] Showing that a "cause" has no effect can occasionally have professional payoff—but it requires a very large sample or high-powered natural experiment to be able to convincingly demonstrate the absence of an effect. *See supra* notes 15-16 and accompanying text. These are rare.

[215] *See supra* notes 58-63 and accompanying text.

[216] *See id.*

[217] Even then, this is usually only when the sample is large enough to be able to reject small or moderate-sized effects—a condition which is not frequently met.

E. *On Research and Knowledge Generation*

This Article makes some big claims. Lurking in the background of these claims are some pretty big questions. To mention just a few: Why aren't my empirical claims more broadly known? Why do so many people hold the engineer's view of social change if it's not supported by the evidence? Shouldn't the academics and policymakers working in this space know better? If research paradigms are so resistant to the knowledge that *they themselves generate*, how can we be confident in our systems of knowledge generation?

These are difficult questions that I will not fully grapple with here. But they are such glaring subtext that I wanted to at least sketch out a few thoughts.

First, the engineer's view is extraordinarily appealing. It would be great if social processes were easily understood and manipulable. It would be fantastic if we could achieve meaningful change with a series of interventions that had been piloted and proven efficacious before scaling up. When a vision is so compelling, it becomes something that people want to be true. Its promise brings people together and helps cross boundaries.[218]

Second, the people best positioned to dispel the myth are those who stand to lose the most from its absence. If the world works in an easily engineerable way, then those trained in empirical causal inference hold an elevated position in the gallery of experts. Their skills are uniquely well-suited to the task at hand: mapping out the causal structure of the social world to show how to improve it. However, if social change doesn't work according to the engineer's view, then it's no longer clear who the relevant experts are. Empirical causal inference no longer holds a special seat at the table.

Third, many researchers don't think of the engineer's view as a myth. Many appear to believe it, or at least accept it in a background way. Why? Maybe because everyone else is acting as if they believe it and so it has become sort of a shared cultural truth. Maybe because it's hard to mentally correct for the distorting influence of researcher incentives, even if these incentives are widely known. When empirical scholars think about research in their field, they tend to think about the studies they've seen presented or that have been published in prestigious journals. These often purport to demonstrate a strong causal relationship between intervention X and outcome Y. Researchers don't see all the studies that were left at the wayside because results were not statistically significant, not-novel, or otherwise unpreferred. Intellectually, people know this. But they may not fully account for it when thinking about the literature.

None of these are fully satisfying answers. Pointing to the biases produced by researcher incentives as an explanation for the persistence of the engineer's view begs the question of why such incentives exist in the first place. It's a circular

[218] Klingele, *supra* note 4, at 562 ("Data, with its promise of impartiality, predictability, and rationality, can be a powerful unifier in modern America, and the rhetoric of evidence-based practice met an especially receptive audience in the world of sentencing and corrections, where decisionmakers have long struggled to avoid decisions about punishment that often feel unanchored or even arbitrary.").

logic: these incentives exist in part because researchers see their project as one of mapping the causal structure of the social world in order to help improve it. In other words, the engineer's view persists because the engineer's view forms the basis of the research paradigm.

Philosophers of science have grappled with such issues for a long time.[219] Science is a human endeavor, a tiny society within society of those grasping toward some semblance of truth. The process is not always as direct as one might hope.

CONCLUSION

Some might see the central claims of this Article as depressing. A world characterized by stabilizing forces that resist change could be seen as a trap, a vortex of inescapable and oppressive social forces. I have a slightly different perspective, one which harks back to an argument presented when discussing the scope of my claim. In an indirect way, this Article celebrates the strength and creativity of the human spirit. The fact that outside forces—interventions—are largely unsuccessful at engineering change in people's lives does not necessarily mean that humans are powerless beings in the throes of social forces. Rather, it suggests that people have already fought to create the best lives they could for themselves given the circumstances. Any barriers to success that were readily moveable had already been moved—by people themselves and their communities. In econ-speak, people had maximized their utility subject to constraints.

That being said, the constraints that remain appear to be deep, structural, and hard to shift. That doesn't mean they are immovable, but just that they usually aren't moveable with the type of intervention evaluable via RCT. As for how to move them—I don't know. Moreover, I don't think we *can* know, or at least not with the high levels of confidence promised by the engineer's view. We will proceed, but must do so with the humility of uncertainty.

[219] *See generally* THOMAS S. KUHN, THE STRUCTURE OF SCIENTIFIC REVOLUTIONS (1962) (arguing progress within science can be limited by scope of current scientific paradigm).