Lessons from the Field:
Conducting Randomized Experiments to Improve Civil Legal Aid Services

Christopher L. Griffin, Jr.
The University of Arizona College of Law

Civil legal aid stakeholders serving litigants in eviction, small claims, debt collection, and family law cases, among others, are calling for evidence-based practices to guide their work. They want to know the strategies lawyers should still pursue and or disregard as no longer effective. Which innovations deliver on their promise, and which fall short? Quantitative social science provides tools for answering these questions with the power of causal inference. And, to that end, empiricists with legal training are embedding evaluations within legal aid operations. This paper describes three prominent methodological examples, with particular attention to the randomized control trial ("RCT") and discusses their application to field evaluations. It provides intuition behind program evaluations and a non-technical primer for the covered research methodologies. The paper also derives key lessons for designing and launching RCTs based on partnerships between researchers and legal aid providers. It concludes with suggestions for fully integrating evidence-based practices and evaluation in civil legal aid operations wherever attorneys commit to improving the effectiveness of their services.

I. INTRODUCTION

The global legal aid community faces daunting challenges in response to the service provision crisis. Legal needs surveys remind us of the myriad human problems that depend on courts and informal process for resolution.[1] Measures of the "justice gap" reinforce how many people try to address their legal problems without adequate assistance.[2] At the same time, governments are pursuing an unprecedentedly ambitious agenda for the future. The chief example is Target 16.3 of the Sustainable Development Goals, through which United Nations member states have pledged to "[p]romote the rule of law at the national and international levels and ensure equal access to justice for all" by 2030.[3] At the very least, these trends seem incompatible. At most, they are entirely irreconcilable. Yet lawyers have embraced this tension to reform the profession with a more user-centered model.

Long-held beliefs about best practices, however, threaten to undermine progress. Take for instance the notion that full legal representation is a necessary condition for resolving most civil legal disputes. The intuition is understandable; how could a litigant possibly do better for herself appearing pro se than with counsel at her side? The same attorney-centric line of thinking supports the claim that non-lawyers cannot (and should not) undertake meaningful roles in the formal dispute resolution process. On this account, only those with a legal education and a license to practice can provide effective support. Both statements, without more, are found wanting. They are *possibly* true yet for decades have never undergone close scrutiny. Rigorous evaluation offers the best opportunity to verify the profession's dominant attitudes toward service provision.

One of the most interesting features of the universal access to justice movement finds legal aid stakeholders doing just that: turning to data for guidance. Services providers frequently invoke "evidence-based practices" as the engine for overdue systemic change.[4] Not all lawyers necessarily conceptualize evidence-based practices

---

[1] *See, e.g.*, ORGANISATION FOR ECONOMIC COOPERATION & DEVELOPMENT, LEGAL NEEDS SURVEYS AND ACCESS TO JUSTICE (2019) (providing tools and recommendations for measuring legal needs); OREGON LAW FOUNDATION, BARRIERS TO JUSTICE: A 2018 STUDY MEASURING THE CIVIL LEGAL NEEDS OF LOW-INCOME OREGONIANS (2018), *available at* https://olf.osbar.org/files/2019/02/Barriers-to-Justice-2018-OR-Civil-Legal-Needs-Study.pdf.

[2] *See, e.g.*, LEGAL SERVICES CORPORATION, THE JUSTICE GAP: MEASURING THE UNMET CIVIL LEGAL NEEDS OF LOW-INCOME AMERICANS (2017).

[3] UNITED NATIONS, DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS, STATISTICS DIVISION, GLOBAL INDICATOR FRAMEWORK FOR THE SUSTAINABLE DEVELOPMENT GOALS AND TARGETS OF THE 2030 AGENDA FOR SUSTAINABLE DEVELOPMENT 18 (2017), *available at* https://unstats.un.org/sdgs/indicators/Global%20Indicator%20Framework%20after%202019%20refinement_Eng.pdf

[4] *See* PEW-MACARTHUR RESULTS FIRST, HOW POLICYMAKERS PRIORITIZE EVIDENCE-BASED PROGRAMS THROUGH LAW: LESSONS FROM WASHINGTON, TENNESSEE, AND OREGON 1 (2017), *available a*t https://www.pewtrusts.org/-/media/assets/2017/04/rf_how_policymakers_prioritize_evidencebased_programs_through_law.pdf ("Evidence-based policymaking, which relies on rigorous analysis of program results to inform budget, policy, and management decisions, is one strategy gaining support among public leaders who want to reduce wasteful spending, expand successful programs, and strengthen accountability."); Michelle N.

the same way. The idea might indicate any approach not based solely on clinical judgment. It could alternatively refer to a new procedure subjected to successful empirical testing. Whatever the definition, evidence-based practices reflect the search for outside verification that long-held professional assumptions work or do not. Legal aid organizations confront financial, human, and other resource constraints but do not always allocate them efficiently in light of those limitations. Evidence-based practices reorient service provision toward the optimal use of time and talent in light of those constraints.

As legal aid moves in this direction—subjecting its service models to empirical examination—advocates and scholars have proposed a multitude of innovative solutions.[5] The proliferation of ideas has even outpaced efforts to demonstrate that any one of them actually improves litigants' experiences. The law is not accustomed to developing the tools of its trade outside of the formal education system, the firm-based organizational model, and other traditional sources. As a result, lawyers have spawned genuine innovation much later than other professionals. Physicians accomplish through clinical research and technologists create through laboratory experimentation what lawyers have just begun to grasp.[6]

Several social science disciplines, including statistics, economics, and psychology, inform evidence-based practices through "program evaluations." Program evaluations test interventions ranging from the discrete and targeted (e.g., providing self-help materials to unrepresented litigants) to the systemic (e.g., civil rights legislation). In plain terms, they ask whether a new policy or procedure measurably affected outcomes deemed important to legal stakeholders. Some program evaluations purport to show that the policy *caused* the changes observed. It is this quest for causal inference that legal scholars are now exporting from the academy into the real world. No other quantitative approach reliably demonstrates whether methods for administering civil legal aid actually deliver (or do not) on their promise.

This paper explains and discusses the state of the art in evaluating access to justice initiatives. It focuses on the randomized control trial, the preeminent study design for demonstrating causal effects. My purpose is less to glorify the methodology than promoting an understanding of what commitment to rigorous evaluation entails for researchers and practitioners. An evaluation in the real world, in real time, forces

---

Meyer, Legal Experimentation: Legal and Ethical Challenges to Evidence-Based Practice in Law, Medicine and Policymaking 2-3 (Nov. 5, 2012) (unpublished manuscript), *available at* https://ssrn.com/abstract=2130828 ("[T]he era of "big data" and information technologies, in turn, promise relatively low-cost ways of evaluating program efficacy and efficiency. And so it is little surprise that momentum is building for continual, sustained evidence- based practice (EBP) within both the executive and legislative branches of federal government . . . .").

[5] *See, e.g.*, JEANNE CHARN & RICHARD ZORZA, CIVIL LEGAL ASSISTANCE FOR ALL AMERICANS (2005).

[6] D. James Greiner & Andrea Matthews, *Randomized Control Trials in the United States Legal Profession*, 12 ANNUAL REVIEW OF LAW & SOCIAL SCIENCE 295, 296 (2016) ("[T]here is no comparing the evidentiary basis for the standard of care as between medical and legal professionals. Practitioners of medicine chose to transform their profession into a science. Practitioners of law did not.").

difficult questions and choices on lawyers. Research can lead to introspection and potentially uncomfortable realizations about suboptimal practices. But that process can also be the small price to pay for larger cost savings and better client experiences. I review the most critical lessons from designing and conducting legal experiments that organizations seeking answers from the data should appreciate. I argue that the profession should embrace an agenda where legal service providers embed rigorous evaluation protocols in new service models before launching them. An evaluation could occur as part of a "research and development" phase or a limited-scope pilot. Whenever and wherever the data analysis takes place, lessons from the field will recursively improve how researchers and practitioners develop a scientific evidence base for civil legal services.

The paper proceeds as follows. Part II first explains legal program evaluation design and then examines two frequently applied methodologies. Section II.B outlines how researchers and field collaborators build randomized field experiments in the law and why this approach dominates other, less suitable statistical methods. Part III reflects on five values from the brief history of civil justice experiments that stakeholders and research collaborators should take into account when pursuing program evaluations. Part IV concludes.

II. THE SEARCH FOR CAUSAL INFERENCE IN THE LAW

The empirical legal studies movement seeks to explain the adjudicatory and transactional outcomes we observe. One might hypothesize, for example, that factor *x* in the justice system causes outcome *y* to occur. But such a claim requires a complete accounting of other factors, legal and otherwise, that also could impact *y*. Otherwise, we might attribute to *x* a causal relationship that truly exists between *y* and *z*. This endeavor has vexed generations of quantitative social scientists, especially those studying justice systems. Legal process depends on complex forces that are often difficult to measure. The analyst might not even know what some of the other factors are.

These truths underlie the problem of "endogeneity," a serious impediment to causal inference.[7] Endogeneity refers to a variety of statistical plagues, including (1) correlation between factors accounted for and not accounted for; and (2) causation running from outcomes to explanatory factors. Researchers have developed solutions to the endogeneity problem, and this Part surveys several common methods that are better and worse on this front. Section II.A introduces two conventional approaches to

[7] *See, e.g.*, Emily S. Taylor Poppe & Jeffrey J. Rachlinski, *Do Lawyers Matter? The Effect of Legal Representation in Civil Disputes*, 43 PEPP. L. REV. 881, 888 (2016) (citing as an endogeneity problem the difficulty in assessing the value of civil representation when representation itself might depend on the plausibility of the claim); Robert Weisberg, *Empirical Criminal Law Scholarship and the Shift to Institutions*, 65 STAN. L. REV. 1371, 1374 (2013) ("The sharp and sustained rise in the rate of incarceration that was simultaneous with the 1990s crime drop has naturally led to empirical research about its causes and cost effectiveness. This research has necessarily faced daunting questions of endogeneity-that is, questions about the relationship between the crime rate and the size of the prison population, or about whether the increase in incarceration is itself the major cause of the crime drop.") (footnote omitted).

statistical inference in the law, and Section II.B concentrates on the randomized control trial. The non-technical discussion is designed to motivate the intuition behind rigorous evaluation and to convince civil legal aid stakeholders to avoid less trustworthy data analysis whenever possible.

Throughout this Part, I use a stylized program evaluation example for context. Imagine a legal aid organization that regularly represents tenant-defendants in eviction proceedings. Its lawyers possess a strong understanding of local rules and customs in the jurisdiction's housing court. The organization's executive director has deduced that staff attorney experience and skill help hundreds of tenants avoid the particularly adverse consequences of displacement (e.g., homelessness) every year, even if the tenant ultimately moves out. An eviction crisis is looming in the jurisdiction. But the organization's budget severely constrains its ability to hire more lawyers and assist even more tenants. What can the executive director do with relatively fixed personnel numbers? A law school clinic has suggested deploying self-help materials, i.e., information that guides users through completion of a specific legal process. But the executive director harbors deep reservations about their usefulness. He questions how the materials could ever positively impact those the office cannot serve directly through representation, however brief. After all, how could a static instruction manual possibly compete with a flesh-and-blood attorney?

A. Empirical Legal Methodologies

1. General Research Design Components

If the executive director in our hypothetical genuinely wants answers to those questions and is open to incorporating evidence-based solutions in the organization's practice model, he could collaborate with empirical researchers. The research team would design a program evaluation focused on the relative benefits of representation offers[8] and self-help material provision. The purpose, as noted above, would be to avoid purely clinical judgments in either direction. Otherwise, the executive director will depend on unproven gut instinct. In all likelihood, the executive director has not conducted an empirical study before; thus, he probably has not identified a precise research question for the study team to pursue. This step is the first that any legal services provider will confront and often requires refinement over multiple conversations.

Crafting the program evaluation question generally involves three elements: a primary outcome measure, an intervention, and a hypothesized relationship. Examples

[8] The distinction between an offer of representation and representation per se is an important one. No ethical study involving human subjects could involve a design where litigants are forced into a particular attorney-client relationship. Individual must always retain their discretion to seek alternative counsel or no counsel at all. This point is emphasized in Greiner & Pattanayak's study of randomized representation offers in a law school clinic. *See* D. James Greiner & Cassandra Wolos Pattanayak, *Randomized Evaluation in Legal Assistance: What Difference Does Representation (Offer and Actual Use) Make?*, 121 YALE L.J. 2118, 2127-28 (2012).

abound in the academic literature.[9] The primary outcome must be the activity, event, or status the organization cares about the most. It must be amenable to quantification, either as a continuous variable (e.g., damages awards, default rates) or a discrete one (e.g., plaintiff win or loss). Settling on a primary outcome is usually not straightforward. Legal aid stakeholders might rank order priorities differently than their colleagues. Some outcomes, such as changes in attitudes and other experiential concepts, resist precise, reliable measurement in the absence of validated instruments on the subject. At the end of the day, the research question isolates the payoff that the organization aims to influence or change. Our legal aid executive director might want to focus on the percentage of cases resulting in a stay of execution. His deputy might favor the defendant appearance rate. Both choices are valid for virtually any program evaluation study design. But the organization must select one as the primary outcome, leaving open the possibility of testing effects on other outcomes (or not).[10]

The intervention, on the other hand, tends to be self-evident. It is the policy or program with unknown potential to improve the outcome selected. In our legal aid example, the intervention is the self-help packet. We study whether, in this case, an information intervention affects the chosen outcome *because* we do not have evidence that they work as predicted. Identifying this component of the evaluation requires little explanation, but the study can succeed or fail if decisions about when and how to apply the intervention are given short shrift. Suffice to say the intervention's form and timing should be a function of ingenuity, responsible risk-taking, and heterodox thinking.

This adventurous spirit dates back in the United States to at least the late 1930s. A Philadelphia attorney "believed it could be demonstrated that there was a place in a large city for law offices which were aimed to serve householders in the lower income group" and that "there was a vast field of preventive law which had scarcely been explored by the lawyer in general practice."[11] In response, he and a handful of other Philadelphia Bar members devised a remarkable plan:

> The basic purpose of this experiment was to determine whether or not the public wished a service which it was not then receiving. We also wanted to test out the practice of preventive law. We knew that the big businessman had been accustomed to consult his lawyer before taking any important step in his affairs, but we suspected that the householder had not. We thought it likely that we would find that the householder usually waited until the necessity for immediate litigation arose before consulting a lawyer. In addition, we wished to learn whether or not a plan of this sort would be helpful to the

[9] *See, e.g.*, Daniel E. Ho, Sam Sherman & Phil Wyman, *Do Checklists Make a Difference? A Natural Experiment from Food Safety* Enforcement, 15 J. EMPIRICAL LEGAL STUD. 242 (2018); Thomas J. Miles. *Does the "Community Prosecution" Strategy Reduce Crime? A Test of Chicago's Experience*, 16 AM. L. & ECON. REV. 117 (2014); Taisu Zhang & Xiaoxue Zhao, *Do Kinship Networks Strengthen Private Property? Evidence from Rural China*, 11 J. EMPIRICAL LEGAL STUD. 505 (2014).

[10] There is nothing wrong a priori in testing a program's effects on more than one outcome measure. The statistics must be adjusted, though, because of what statisticians call "multiple testing." See Joseph P. Romano, Azeem M. Shaikh & Michael Wolf, *Hypothesis Testing in Econometrics*, 2 ANN. REV. ECON. 75 (2010).

[11] Robert D. Abrahams, *The Neighborhood Law Office Experiment*, 9 U. CHI. L. REV. 406, 406 (1942).

> economics of the legal profession, particularly in aiding young lawyers to obtain a practice.[12]

Abrahams' idea exemplifies a well-designed intervention: easy to explain and skeptical of the status quo. Local, informal "clinics" were at once obvious innovations and wholly shocking to the bar. The proposal might strike us as quaint now, but attorneys at the time dismissed or condemned the intervention. In fact, criticism leveled against contemporary innovations—courthouse navigators,[13] limited license legal technicians,[14] and online dispute resolution platforms,[15] among others—sounds uncannily similar. "[A]ny change in time-honored methods of practice tends to lower the 'dignity of the profession' . . . ."[16] "Your idea is a good one on paper, but, like everything else in our profession, nothing will be done about it."[17] Developing evidence-based practices begins with a creative spark to develop a counterintuitive solution and culminates in an empirical demonstration of effectiveness.

Finally, the Philadelphia neighborhood law office experiment implied a clear hypothesis, the presumed relationship between the intervention and the outcome. Hypotheses can be "one-" or "two-sided,"[18] but they must be consciously held. Adherence to a hypothesis does not necessarily mean that the stakeholder and researcher are biased toward one outcome. The expected correlation provides a baseline against which they will compare the observed data. In our eviction representation example, therefore, the executive director could appropriately conjecture that the self-help materials will perform worse than full attorney representation. He could just as well suggest that the packet's impact on the primary outcome will be statistically indistinguishable from representation. But he should explicitly formulate a hypothesis.

---

[12] *Id.* at 408.

[13] *See* REBECCA L. SANDEFUR & THOMAS M. CLARKE, ROLES BEYOND LAWYERS: SUMMARY, RECOMMENDATIONS AND RESEARCH REPORT OF AN EVALUATION OF THE NEW YORK CITY COURT NAVIGATORS PROGRAM AND ITS THREE PILOT PROJECTS (2016), *available at* http://www.americanbarfoundation.org/uploads/cms/documents/new_york_city_court_navigators_report_final_with_final_links_december_2016.pdf

[14] *See* Patrick McGlone, *Can Licensed Legal Paraprofessionals Narrow the Access-to-Justice Gap?*, ABA JOURNAL (Sept. 6, 2018), http://www.abajournal.com/news/article/can_licensed_legal_paraprofessionals_narrow_the_access_to_justice_gap.

[15] *See* Erika Rickard, Pew Charitable Trusts, Online Dispute Resolution Offers a New Way to Access Local Courts, *available at* https://www.pewtrusts.org/en/research-and-analysis/fact-sheets/2019/01/online-dispute-resolution-offers-a-new-way-to-access-local-courts.

[16] *Id.* at 407.

[17] *Id.*

[18] *See* Lyle V. Jones, *Tests of Hypotheses: One-Sided* vs. *Two-Sided Alternatives*, 49 PSYCHOLOGICAL BULLETIN 43, 44 (1952) ("More often than not . . . our hypotheses have a directional character. We are interested in whether or not a given [intervention] improves . . . performance . . . . The appropriate experimental test is one which takes this into account, a test of the null hypothesis against a one-sided alternative.").

2. Select Methodologies

Once stakeholders select the core study elements, the researcher considers which evaluation methodology to use. Much of this decision process, involving data structures and details about the service provision under evaluation, lies beyond the scope of this paper. Instead, I summarize two common, but less rigorous, approaches to understanding whether innovations work. Section II.B follows with an explanation of randomized control trials.

i. Before-After Studies

Experience shows that civil justice stakeholders gravitate toward and, as a matter of policy, prefer the most intuitive evaluation methodology: before-after comparisons. As the name suggests, this study design (essentially) compares the mean of the outcome measure for some time period before the intervention and the mean outcome value for some time period after.[19] The measures can be compared with or without accounting for other possible explanations.[20] The preference for before-after studies follows from their relatively easy implementation. The legal aid stakeholder and research team only ensure that the intervention is deployed on a given date and remains in constant use until the study is over.

I graphically depict in Figure 1 data that a before-after study could produce. The figure plots values of an outcome measure on the vertical axis against time on the horizontal one. A second vertical line represents when the organization introduced the intervention, here the self-help materials. One assumption embedded in the graph is that *no* litigants received self-help materials before the time marked by the vertical line, and *all* litigants received them afterward. The legal aid organization would not conceivably provide services this way, unless it counterfactually had no attorneys committed to full eviction representation before the vertical line and then only provided the self-help materials. (This design would evaluate self-help materials' effectiveness against a baseline of no assistance with eviction cases.)

[19] One of the most well-known (and later criticized) before-after empirical legal studies about a legal issue is Isaac Ehrlich, The Deterrent Effect of Capital Punishment: A Question of Life and Death, 65 Am. Econ. Rev. 397 (1975). As Donohue and Wolfers point out, "[e]ven though Ehrlich's 1975 study was to be later discredited, the real problem was not that a flawed empirical paper had been written, but rather that there were those who leapt to use it as a tool to advance the goal of reinstating capital punishment in the United States before the validity and reliability of the work had been fully explored." John J. Donohue & Justin Wolfers, *Uses and Abuses of Empirical Evidence in the Death Penalty Debate*, 58 STAN. L. REV. 791, 844 (2005)

[20] Accounting for alternative explanations corresponds to what empiricists call "controlling for" the effect of other variables. Most before-after studies include controls to separate their effects from the intervention's impact.

Figure 1: Before-after Study

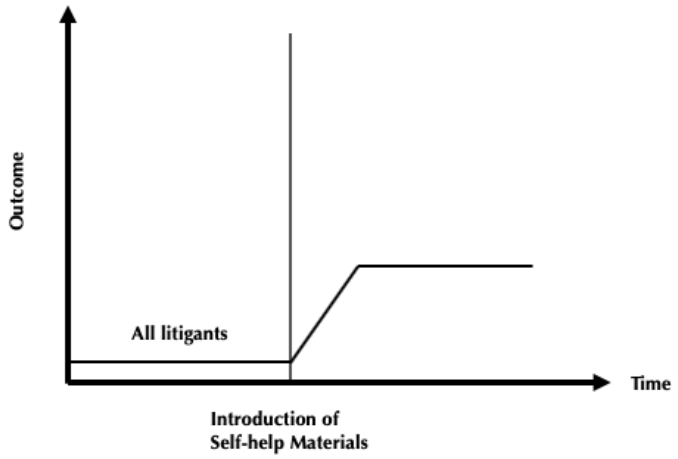


Figure 1 suggests that tenant outcomes, which were uniformly low before the self-help materials arrived, improved linearly after the organization deployed them and then plateaued after a short while at a new, higher level. The mean outcome value on the right side of the vertical line unequivocally exceeds the mean value to the left. The legal aid organization would seemingly have a basis in evidence that the innovation outperformed whatever assistance was provided during the "before" period in terms of client outcomes.

The severe limitations of this study should be obvious. A clear flaw is that this analysis does not account for some other factor that emerged at the same time as the innovation's introduction and improved outcomes. What if the housing court launched a mediation program available to all tenants around the same time that the legal aid organization unveiled its self-help materials? The chance to negotiate in a formal mediation setting and achieve a more agreeable settlement for both the landlord and tenant could have been just as responsible for the better outcomes. Even if the researcher accounts statistically for the mediation program, other factors could have affected tenant outcomes at the same time as the self-help packets that cannot be measured or remain unobservable. If so, the before-after study design falls far short of the standard for causal inference.

Another clear problem with the before-after approach is that the composition of litigants in the "before" period could differ substantially, if not entirely, from the litigant population in the "after" period. When the population changes enough, any differences in the motivations, litigation strategies, or other characteristics between the two groups will confound the study. Researchers call the complete set of before-after study limitations threats to "internal validity."[21] Under these conditions, the researcher will

[21] For a helpful overview of internal validity constraints on before-after studies, see LYNDA S. ROBSON ET AL., CENTERS FOR DISEASE CONTROL AND PREVENTION, GUIDE TO EVALUATING THE EFFECTIVENESS OF STRATEGIES FOR PREVENTING WORK INJURIES: HOW TO SHOW WHETHER A SAFETY INTERVENTION REALLY WORKS 19-27 (2001), *available at* https://www.cdc.gov/niosh/docs/2001-119/pdfs/2001-119.pdf.

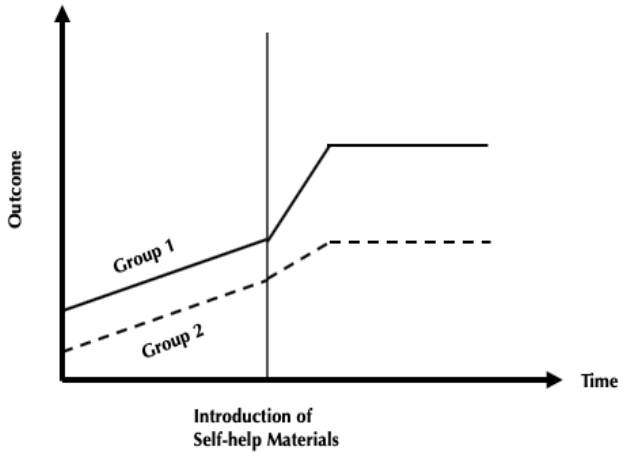
struggle to obtain reliable causal inference because she could not isolate the effect of the intervention alone. Too many other factors could plausibly explain the result as well as, or better than, the innovation.

ii. Quasi-Experimental Methods

The abundance of caution when interpreting before-after studies motivates the need for comparison, or "control," groups. Take a variation on the problem identified above in which another factor besides the intervention affects all litigants. If the sample of litigants remains exactly the same before and after the launch date, everyone, by definition, receives exposure to the intervention. Because program evaluations involve comparisons between persons who are and are not subject to an intervention, a group must exist that had no access to the intervention. Thus, studies that compare outcomes between two periods of time lack a crucial element. They do not have a counterfactual set of litigants proceeding on the same timeline who were not exposed to the new program or policy.

In response, social scientists "borrow[ed] a page from medical studies" to incorporate what are known as "control" observations.[22] These comparison observations resemble the formal control subjects in a clinical trial, while those who receive access to the intervention are called "treated." The researcher tracks and compares the outcome trends for the treated and control groups using the same average values from the before-after design. The difference lies in the presence of suitable comparators, which did not exist in Figure 1 (where every litigant was provided the self-help materials). Such a quasi-experimental design appears graphically in Figure 2. Now the executive director can do more than compare results when everyone has self-help materials to a world when no one has them. He can also compare those two situations for two sets of litigants: those who always received or never received the materials.

Figure 2: Quasi-Experimental Study



[22] Donohue & Wolfers, *supra* note 19, at 798.

Figure 2 shows one clear similarity between Groups 1 and 2. The upward trend in outcomes was identical for both sets of litigants; the slopes of the Group 1 and Group 2 curves were equal. The main difference between them is that Group 1's outcomes were better than Group 2's (and constantly so) before the intervention launched. In plain terms, both sets of litigants appear to have benefitted from some unknown factor(s) over the pre-intervention period, but Group 1 was always in better position than Group 2 during the same time. Such pre-intervention patterns lend themselves to quasi-experimental analysis even though the two groups were significantly different before the study. The statistical analysis depends exclusively on whether and in what direction the two trends shift *after* the intervention is introduced. If the two groups were already on divergent trajectories, i.e., the slopes of the two curves were not the same, before the program launched, one suspects that the groups were already subject to dissimilar forces. It becomes exponentially more difficult to isolate the causal effect of an intervention when this is true. But if the trends follow the same path in the "before" period, i.e., if the slopes of the curves are the same, one can more reliably attribute any "after" period slope changes to the intervention.[23]

Figure 2 depicts the intuition behind so-called "difference-in-differences" ("DID") studies. The graph, at first glance, shows only one difference—the *between-group* one measured as the distance between the Group 1 and Group 2 curves. But it also contains a second difference that compares the *within-group* change before and after the intervention. The DID approach performs this operation first; the researcher effectively asks what a before-after study reveals separately for the treated and control groups. If Group 1 received the self-help materials, it seems that its members had better experiences after the office introduced self-help materials. But Group 2 individuals also saw an uptick, albeit a slight one, in their average outcomes. How does this result comport with the study design?

An answer is found in the second difference, computed as the difference between Group 1's before-after change and Group 2's. The DID methodology does not rule out benefits (or disadvantages) accruing to the control group. Instead, it accounts for them in the first difference. Consider an alternative version of Figure 2 in which the slope of the Group 2 curve in the "after" period is exactly the same as Group 1's. The before-after difference in mean outcome values (as a percentage change) would be identical between the two sets of litigants. Therefore, the difference in differences between Groups 1 and 2 would be zero. Stated otherwise, something *other than the intervention* most likely affected both groups and caused them to experience the same improvements. We cannot ascribe those changes to the intervention, because the Group 2 subjects by construction did not encounter it.

[23] Any pre-existing, consistent differences in outcomes during the "before" period will not affect interpretation. The study would be even easier to explain if Groups 1 and 2 shared the exact same pre-intervention trend, i.e., if there were only one curve to the left of the vertical line representing outcomes for both Groups 1 and 2. But such uniformity is not required, only that the slopes of the two curves are identical.

Returning to the actual Figure 2, we nevertheless observe a slight upward trend in Group 2's outcome. This result could have been driven by a new rule change (e.g., an extension of time between service of the eviction complaint and the hearing).[24] Group 1 members presumably benefitted from the same adjustment, but the percentage increase in outcome values is clearly greater than in Group 2; the slope of the Group 1 curve is now higher than Group 2's just after the intervention is introduced. The researcher could infer that this additional increase was a function of the intervention, the self-help materials. Such inference would only be proper if the two groups were balanced, or identical on observable characteristics, in the "before" period. In other words, the researcher needs to show that she identified treated and control groups whose features (but not necessarily their outcomes) were equivalent before the intervention launched. Even the best quasi-experimental studies cannot meet this standard and thus do not yield valid causal inference.

B. Randomized Control Trials Explained

The basic concept behind a randomized control trial ("RCT")—also known as a randomized field experiment—is straightforward. The researcher artificially creates two groups of subjects, in our example, the potential eviction clients. Conversely, researchers look to other jurisdictions or organizations for legitimate comparison observations in most quasi-experimental studies Just as in the quasi-experimental design, though, we call them treated (exposed to the intervention) and the control (not exposed to the intervention) groups. The main differences between the RCT and the quasi-experiment are twofold. First, most RCTs randomize units (in our hypothetical, eviction litigants) into the two groups. Each person effectively enters a lottery in which there should be an equal chance of receiving the intervention versus not. Quasi-experiments typically provide the intervention "dosage" to selected clusters of people.[25] If the researcher can collaborate across provider offices, she might choose one legal aid organization to have all its potential clients receive self-help materials and the other office to only provide full representation. If the researcher did not assign the offices randomly to treatment and control, the experiment would not be an RCT. Even if she did, the RCT would be unorthodox and probably not very helpful. A true RCT more likely would take place within just one office and would treat each litigant as a separate potential member of the treated and control groups.

[24] The relevance of this rule change would be that more time between complaint receipt and hearing provides tenants with more opportunities to gather evidence and witnesses in preparation for the hearing.

[25] Quasi-experiments that arise because of uncoordinated changes to law or policy across jurisdictions are known as natural experiments. Many contemporary articles in empirical law and economics exploit natural experiments to approach, but not achieve, causal inference. *See, e.g.*, Oren Gazal-Ayal and Raanan Sulitzeanu-Kenan, *Let My People Go: Ethnic In-Group Bias in Judicial Decisions—Evidence from a Randomized Natural Experiment*, 7 J. EMPIRICAL LEGAL STUD. 403 (2010); Daniel E. Ho & Mark G. Kelman, *Does Class Size Affect the Gender Gap? A Natural Experiment in Law*, 43 J. LEGAL STUD. 291 (2014); Dara Lee Luca, *Do Traffic Tickets Reduce Motor Vehicle Accidents? Evidence from a Natural Experiment*, 34 J. POL'Y ANALYSIS & MGMT. 85 (2015).

Randomization methods theoretically range from the proverbial coin flip to complex algorithms.[26] Some studies rely on an immutable characteristic, easily subject to independent monitoring, such as a client's month of birth or a case number randomly generated by a computer system in the clerk of court's office. Whatever the method or identifier used, it must convince social scientists that membership in the treated and control groups was uncorrelated with any characteristic that could credibly affect the outcome, too. In other words, the randomization scheme must rule out the problem of endogeneity noted above. The sorting of units into experimental conditions theoretically ensures balance between participant groups for all conceivable observable (and unobservable) factors other than the intervention that could influence the outcome. And randomization scheme ideally will be easy to implement by people outside of the research team. Many legal RCTs increasingly rely on legal aid stakeholders to follow a randomization protocol because the research team cannot be on site or on call to assign each unit in real time.

Before study units, i.e., participants, receive experimental assignments, the research team and stakeholders specify what the treated and control conditions entail. Technically, the randomization process need not yield only two *treatment arms*; an RCT can include two, three, or more treatments compared to a control.[27] In some situations, the researcher might prefer to compare the intervention to a truly "no-intervention" control condition.[28] These designs offer nothing to study participants; the control group members are both excluded from the intervention and are not provided any other meaningful service or information.

A recent, high-profile example of a (nearly) no-intervention control study with significant consequences for interpreting results is Greiner & Pattanayak's evaluation of representation offers at a Harvard Law School clinic.[29] The treated condition was an offer of representation from the clinic, and the control condition was no offer but provision of the "names and telephone numbers of other legal services providers in the area" that might be able to offer representation.[30] Some might consider this control condition as including an "intervention"—the provision of other lawyers' contact information. Others might understandably view the phone numbers as a perfunctory offer. Because it would be highly unethical for the study to prevent those assigned to the

[26] *See, e.g.*, Ellen Degnan et al., Trapped in Marriage at 24 (unpublished manuscript) (Nov. 28, 2018) ("Our randomization scheme was simple. We created blocks of 10-20 observations and programmed a computer to allocate randomly 0's and 1's within each block."); D. James Greiner & Andrea J. Matthews, The Problem of Default, Part I at 26 (unpublished manuscript) (June 16, 2015), *available at* https://ssrn.com/abstract=2622140; (In the first two weeks of our study, we used a computer to create a group of 24 cases, eight each randomly distributed to our three treatment arms: Control, Limited, and Maximal . . . .").

[27] Expanding the number of treatment arms depends on the "power" of the study, which I discuss in the next Part.

[28] *See, e.g.*, Patrick Pössel et al., *A Randomized Controlled Trial of a Cognitive-Behavioral Program for the Prevention of Depression in Adolescents Compared to Nonspecific and No-Intervention Control Conditions*, 60 J. COUNSELING PSYCH. 432, 433 (2013) (observing that "[m]ost depression prevention studies in adolescents have compared a specific intervention to a no-intervention . . . control").

[29] Greiner & Pattanayak, *supra* note 8.

[30] *Id.* at 2143.

control condition from obtaining a lawyer, some control participants did just that. In fact, "[s]ome claimants randomized not to receive an HLAB offer obtained representation from other service providers, such as private attorneys, Greater Boston Legal Services, the Volunteer Lawyers Project, or the clinical program at Northeastern Law School."[31] Whether to include any service in the control condition is a matter for stakeholders to decide. That choice will alter the interpretation of results and also affect whether intervention cross-over (as in the Greiner & Pattanayak study) is more or less likely to occur.

After the research team and stakeholders set the contours of the experimental conditions, and before launching the study, an empiricist should conduct a *power analysis*. Power represents the researcher's best estimation of how many subjects she needs must enroll to yield valid causal inference.[32] As we saw earlier from the legal aid executive director's hypotheses, the self-help materials could result in three scenarios: improvement relative to representation, worsening relative to representation, and no difference. The latter outcome is the most concerning. Such "null results" might be statistical truths; no causal relationship existed between the intervention and the outcome relative to the status quo. Or the lack of a difference could be a statistical artifact. Perhaps the intervention condition actually is more effective than the control one. If the study does not test a statistically sufficient number of participants, it might not detect the intervention's effect, *even though it exists*. The evaluation would have been "under-powered." To avoid this result, the researcher aims to specify a minimum sample size that will give her the best chance—but no guarantee—that the study will detect true intervention effects, if they exist.

The final pieces of the power analysis puzzle is a *minimum effect size* ("MES") and baseline outcome measures. The empiricist will ask legal stakeholders to specify a threshold difference in outcome values between the treated and control group that they consider relevant and indicative of success. Where possible, they will also estimate outcome values for past litigants, perhaps former clients of the organization, to serve as the baseline. Sometimes, the MES comes from other studies of a similar intervention, and the stakeholder would agree that these prior estimates are floors for measuring success. Or the MES might be a function of clinical judgment and program ambition. The value chosen should balance reasonability with sincerely held predictions. The reason is that, all else equal, smaller MES values require larger sample sizes in order to detect them. The baseline outcome estimate interacts with the MES, too. The closer pre-existing rates for an outcome are to 50%, the larger the sample size will need to be. For example, if the best guess as to the intervention's effect is a 2-percentage-point increase from a baseline of 40%, the required sample size to detect the change will be greater than a 15-percentage-point increase from a baseline of 2%. In short, smaller

[31] *Id.* at 2166-67 (footnotes omitted).

[32] *See* John M. Lachin, *Introduction to Sample Size Determination and Power Analysis for Clinical Trials*, 2 CONTROLLED CLINICAL TRIALS 93, 95 (1981); *see also* John M. Hoenig & Dennis M. Heisy, *The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis*, 55 AM. STATISTICIAN 19, 23 (2001) ("Power calculations tell us how well we might be able to characterize nature in the future given a particular state and statistical study design, but they cannot use information in the data to tell us about the likely states of nature.").

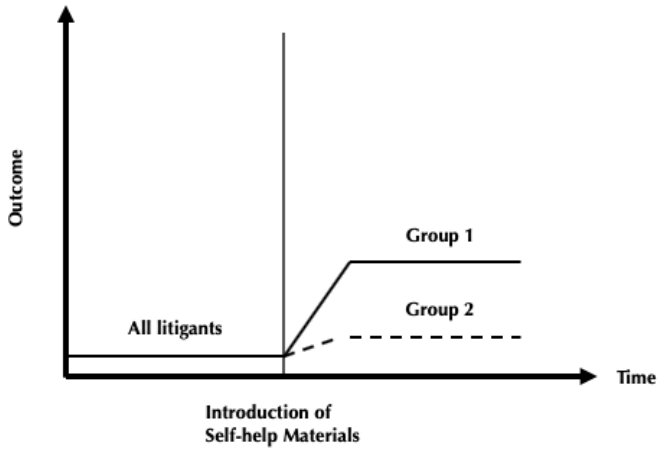
effect sizes are always harder to detect statistically; they are the proverbial needles in data haystacks.

With all of the foregoing elements in place, the RCT program evaluation is ready to launch. To be sure, this presentation substantially simplifies and condenses the process of RCT design. I have described it as such, though, to warn that the instinctive appeal and methodological benefits of a randomized field experiment still demand value judgments from both researchers and legal services providers. The art of RCT construction indeed rivals the science needed to complete the study.

Why go to all of this trouble? Causal inference has long eluded social scientists who use before-after and even quasi-experimental methods. Studies of these varieties usually take naturally occurring data and seek patterns that explain their movement. When a government or organization introduces a new policy, it does so (usually) in a uniform manner. All eligible persons in the jurisdiction or all clients will be treated identically under the new approach. Any researcher hoping to evaluate the policy is hamstrung by the lack of a comparison group. Intuitively, the problem is that there are no pure observational units whose outcomes can be contrasted against the affected population. Identifying a plausible control group, though, does not guarantee causal inference without more. Unobserved or unmeasurable factors will still linger in the background.

RCTs are the only evaluations that take a pre-existing population and separate its members into experimental groups *at the same time the intervention starts.* If the intervention matters it should be apparent in a simple comparison of relative outcomes without controlling for other factors, thereby requiring fewer assumptions about other possible explanations for the same outcomes. Figure 3 provides the graphical intuition for RCTs. Group 1 is once again the treated set, and Group 2 is the control. Notice that, even with an RCT, the control group might still experience an improvement in outcomes. That increase, as measured by the higher slope of the Group 2 curve, is decidedly less than the increase for Group 1. Because assignment to the experimental arms followed a randomized protocol, the Figure 3 between-group difference is the only relevant one. And it is more reliably a direct function of the intervention rather than some other factor(s).

Figure 3: Randomized Control Trial



Properly designed RCTs approximate or achieve the statistical rigor necessary for demonstrating that an intervention caused certain outcomes without imposing onerous assumptions on the study. As a result, one often hears that the RCT is the "gold standard" in program evaluation. The argument over whether this label should go the way of its namesake in 20th Century monetary policy is lively and too voluminous to cover fully in this space.[33] The truth remains that RCTs, when conceived carefully, also require care in analyzing and interpreting the resulting data. Randomized assignment to the intervention is not a panacea for program evaluation success. It is only the beginning, and the next Part reviews some of the primary reasons why.

III. ADVICE FOR LEGAL PROFESSIONALS CONSIDERING A RANDOMIZED FIELD EXPERIMENT

Primers on program evaluation methods, including RCTs, historically appear more often in medical, psychological, and educational studies than in the law.[34] As a result, research in those fields have benefitted from decades of technical and experiential lessons. Meanwhile, the legal profession is still experimenting with experimental studies. The shorter history of RCTs in the law still offers researchers and legal aid stakeholders suggestions for better practices. The practical lessons gleaned from applying RCT methodologies are just as important as mastering the statistical details.

---

[33] *Compare* Pat Dugard, *Randomization Tests: A New Gold Standard?*, 3 J. CONTEXTUAL BEHAV. SCI. 65, 68 (2014) ("Far from regarding classical tests as the gold standard for statistical inference, and randomization tests as a slightly eccentric approach suitable only when we have a single case or small group of participants, it is the randomization tests which may provide the gold standard in the future."), *with* Gareth Parry & Maxine Power, *To RCT or Not to RCT? The Ongoing Saga of Randomised Trials in Quality Improvement*, BNJ QUALITY & SAFETY
221, 222 (2015) (noting that RCTs abstract too much from the questions of how or why an intervention succeeds in the service of generalizable solutions).

[34] *See* Greiner & Matthews, *supra* note 6.

This Part covers five topics that empirical legal researchers often encounter when designing and conducting RCTs. The list is far from exhaustive, and it concentrates on those areas where evaluation requirements tend to conflict with the standard operating procedures of legal aid stakeholders. In the sections that follow, I discuss (1) criteria for deciding when to use RCTs; (2) the ethics of randomization, particularly in legal contexts; (3) the related requirements of Institutional Board Review oversight and informed consent; (4) responding to dissent among stakeholders; and (5) being open to the possibility that the evaluation will not affirm stakeholders' preferred outcome.

A. Prioritizing Legal Aid Evaluation Needs

Not all program evaluations are created equal in the eyes of the researcher. Although Part II made the case for RCTs as the best methodological choice for demonstrating what works, i.e., for approaching causal inference, RCTs are nether appropriate nor necessary for all civil legal studies. One reason concerns the ethical feasibility of randomizing an intervention at all or against a no-intervention control. A health policy program that provides known, essential health and wellbeing services, for example, should not be apportioned via lottery. In other cases, ethical reviewers will insist that the control units receive some form of assistance rather than denial and dismissal. If the stakeholder wants to compare the intervention to a baseline of no assistance, an RCT would be precluded. In addition, some questions on practitioners' minds cannot be answered by RCTs or cannot justify the upfront cost in time and money to conduct. When those questions relate to how the intervention works, the randomized experiment will be of no use. RCTs reveal only whether one program, idea, or policy causes better outcomes; they are silent as to why.[35]

Field RCTs, whether multi-year operations or short-term experiments, always present logistical challenges. Depending on the scope of the RCT and the size of the intervention, stakeholders and/or researchers might need to apply for grant funding. The upfront costs of grant applications with an uncertain probability of success can be enough to frustrate the project. But money is rarely the tightest constraint. Randomized program evaluations in the real world must disrupt or at least alter some process that the legal services organization has followed before the study. If they do not, the research is not testing a true intervention. Standard operating procedures, especially among lawyers, are sticky, and, like the profession writ large, not amenable to sudden change.

[35] *See* Martyn Hammersley, *What Is Evidence for Evidence-Based Practice?*, *in* EDUCATION SCIENCE: CRITICAL PERSPECTIVES 101, 102 (Ralf St. Clair, ed. 2009) ("In the context of medicine, [an RCT] . . . can tell us about survival rates following some treatment, but it may not tell us about survival rates for different categories of person, or about the distribution of side effects across those categories. So, even in the part of medicine where RCTs are most effective, they will not always give us answers to all the relevant questions, despite their great value in answering some."); *see also* Yudhijit Bhattacharjee, *Can Randomized Trials Answer The Question of What Works?*, 307 SCIENCE 1861, 1862 (2005) (quoting criticism that the Department of Education for "plan[s] randomized studies without knowing why an intervention seems to work").

Legal aid stakeholders should therefore consider a spectrum of evaluation questions during conversations within the organization and among potential research collaborators. Issues on the periphery of one's practice should be investigated using less intensive methods. For example, understanding what approaches improve *lawyers'* experiential outcomes and work attitudes would be a legitimate inquiry. But the relative costs of an RCT surely will outweigh any benefits that accrue to the legal aid organization. A well-designed survey administered before and after the intervention will usually suffice. An RCT will not be able to solve the inherent difficulty of precisely measuring attorneys' feelings. The same principle is true for studying practices that are ancillary to the mission of the organization (e.g., whether software tools used in the office are effective) but not for discrete, core features of service provision (e.g., intake and triage procedures).

We also know that, if the organization is more interested in understanding how interventions make a difference, an RCT is not the preferred methodology. Rather, it should pursue a *process evaluation* only, or initiate a process evaluation before designing a fully randomized study. Process evaluations address the following objectives, among others:

> [They] typically examine aspects related to delivery and implementation processes such as fidelity (that is, was it delivered as planned?) . . . and reach. Process evaluations are also concerned with how an intervention has an effect on participants, organisations, and communities, including their response to the intervention and its influence on determinants of outcomes (for example, did it change the identified negative attitudes, communication skills, or community engagement?).[36]

As one team of social scientists has observed, "many RCTs would be enhanced by an integral process evaluation. The additional costs (such as collecting and analysing qualitative data) would probably be balanced by greater explanatory power and understanding of the generalisability of the intervention."[37] In other words, resort to an RCT may never be timely there is no accompanying theory for why the intervention would work. Without that background information, the program evaluation could be useful only in the specific context of one stakeholder's process. Process evaluations, by explaining (and not just revealing) causal mechanisms provide templates for other organizations interested in replicating successful interventions.

---

[36] PUB. HEALTH ENGLAND, GUIDANCE: PROCESS EVALUATION (2018), *available at* https://www.gov.uk/government/publications/evaluation-in-health-and-well-being-overview/process-evaluation.

[37] Ann Oakley et al., *Process Evaluation in Randomised Controlled Trials of Complex Interventions*, 332 BRITISH MED. J. 413, 415 (2006); *see also* Tracy W. Harachi et al., *Opening the Black Box: Using Process Evaluation Measures to Assess Implementation and Theory Building*, 27 AM. J. COMMUNITY PSYCH. 711, 713 (1999) ("[T]heoretical context elucidates explanation or interpretation of the mechanisms through which the intervention effects occur. A program evaluation enhances its utility by examining the theoretical basis of the program and the intervening and contextual factors that mediate the relationship between the program and the ultimate outcome.").

B. Ethical Issues

Legal scholars have not yet developed an ethics of randomization to match their counterparts in medical research.[38] Ideally, it would include generalizable principles that apply across jurisdictions, legal subject matter areas, and intervention types. As that work continues, the field has borrowed ideas from the medical ethics literature. That scholarship in addition to experience and reflection suggest two criteria that should guide most, if not all, legal RCTs: *scarcity* and *equipoise*.

When there are not enough resources to serve every eligible client, which is almost always the case in legal scenarios, randomization is an ethical way to allocate what little there are, particularly if doing so will allow us to learn what works and what does not. The reason is straightforward. Imagine, for example, that not all potential clients will receive full representation from a legal aid organization. The legal aid organization could ethically use a lottery to decide who will be represented when not all can be served, and in doing so, learn how much of a difference representation makes. Scarcity requires some rule for allocation. Legal services will be deployed; we just have to decide how. And randomization is one ethical method that treats people fairly while allowing us to generate knowledge that will benefit others.

The second reason to randomize is based on the concept of equipoise, which was popularized by Professor Charles Fried.[39] Equipoise means that we do not know the answer to a research question already—specifically we do not already know whether some proposed intervention is effective (or harmful). If a researcher knows that a new procedure is harmful or beneficial, there is no reason to study it. As one research ethics scholar frames the dilemma, "[e]quipoise is an ethically necessary condition in all cases of clinical research. . . . Theoretical equipoise exists when, overall, the evidence on behalf of two alternative treatment regimes is exactly balanced."[40] But in law, which has no established tradition of conducting rigorous, evaluations to produce credible evidence of effectiveness, we rarely know at the outset what new ideas will be effective. We have policy preferences, or professional judgments, or educated guesses, but not evidence.

Before the randomized trial launches, one can also be in a state of clinical equipoise, meaning there "is no consensus within the expert clinical community about the comparative merits of the alternatives to be tested."[41] If the program evaluation shows that the intervention is more beneficial than the control alternative, clinical equipoise is disturbed. Clinical equipoise therefore is its own worst enemy, as it were. It

[38] *See, e.g.*, Bernard A Foëx, *The Ethics of Clinical Trials*, 10 ANESTHESIA AND INTENSIVE CARE MED. 98 (2008); Benjamin Freedman, *Equipoise and the Ethics of Clinical Research*, 317 NEW ENG. J. MED. 141 (1987); Richard M. Royall, *Ethics and Statistics in Randomized Clinical Trials*, 6 STAT. SCI. 52 (1991).
[39] *See* CHARLES FRIED, MEDICAL EXPERIMENTATION: PERSONAL INTEGRITY AND SOCIAL POLICY (1974).
[40] Freedman, *supra* note 38, at 143.
[41] *Id.* at 144.

is a prerequisite for conducting a randomized experiment, but it will disappear as soon as the trial produces results favoring the intervention or the alternative.

A well-designed, ethically sound RCT arguably does not require satisfaction of both the scarcity and equipoise conditions. Situations where the former clearly dominates (e.g., when the intervention is an offer of full legal representation) usually permit randomized evaluation so long as the research team is ignorant of the relative benefits of alternative treatment paths. This means that the empiricist will proceed when there is *no* evidence for either version of the hypothesis and when there will not be enough of the intervention to provide to all qualifying participants. On the other hand, viewing alternative policies through the lens of theoretical or clinical equipoise might be a weak basis for RCT evaluation when the stakeholder is now constrained by scarcity. The best argument for an equipoise-based randomized experiment without true scarcity is an overwhelming public policy interest in learning more about the intervention and when randomization does not violate any other legal principle (e.g., due process or equal protection). Current examples are the multiple evaluations of the Public Safety Assessment, an actuarial risk assessment for use at bail hearings.[42]

C. Institutional Review Boards and Consent Procedures

Just as legal services provision is regulated by bar associations and government actors, academic research is policed for compliance with ethical standards. Institutional Review Boards ("IRBs") exist at higher education institutions and some independent organizations to perform this function.[43] Even if the legal stakeholder faces no external obligation or other oversight, any university-based researcher conducting research on human subject data must submit a protocol to its institution's IRB. This requirement extends to analysis of secondary data about persons, i.e., publicly collected, potentially identifiable data; these studies are usually deemed exempt from IRB oversight after submission. Federal law obliges colleges and universities in the United States, for example, to administer IRB supervision of human subjects research to prevent horrific experiments on people from occurring again.[44]

---

[42] For more information, see The A2J Lab, Pretrial Release, https://a2jlab.org/current-projects/signature-studies/pretrial-release/ (last visited Apr. 7, 2019).

[43] *See, e.g.*, Robert J. Amdur & Chuck Biddle, *Institutional Review Board Approval and Publication of Human Research Results*, 277 JAMA 909 (1997); Ralph L. Rosnow et al., *The Institutional Review Board as a Mirror of Scientific an Ethical Standards*, 48 AM. PSYCHOLOGIST 821, 822 (1993) ("A central responsibility of IRBs is to ensure that the potential benefits to the individual research participants (and to society) will be greater than any risks that may be encountered by participation in the research.") (internal citation omitted); Harold Y. Vanderpool, *An Ethics Primer for Institutional Review Boards* 3-8, *in* Institutional Review Board: Management and Function (Elizabeth A. Bankert & Robert J. Amdur, eds. 2d ed. 2006) .

[44] *See* Susan M Reverby, *Listening to Narratives from the Tuskegee Syphilis Study*, 377 LANCET 1646, 1646 (2011) (describing the infamous Tuskegee Syphilis Study and how its ignominious legacy "hung over the heads of those on the government-sponsored *Belmont Report* that articulated key principles for bioethics and regulation for research: respect for autonomy, protection of the vulnerable, beneficence, informed consent, and a promise of justice").

IRBs manage multiple ethical objectives in any one study, but the imperative of informed consent ranks near the top of its priorities. For research that involves more than minimal risk to participants, the institution usually expects signed consent forms from recruited subjects. The process cannot be cursory. The IRB expects the research team in most cases to describe the reasons why the person has been recruited, what enrollment will require of the person, and what benefits and risks she can expect. Providing study details in plain but thorough language—and offering to answer questions—ensures that the recruited person makes a real choice rather than submit to research blindly.

One might not expect licensed attorneys to object to the informed consent principle, and they do not, generally speaking. The strictures of IRB protocols for consent, however, often rankle stakeholder research partners. Researchers and legal service providers must learn to integrate their respective needs as study development proceeds, and experience demonstrates that the consent process dominates the conversation. High-volume, high-intensity service areas such as eviction defense and domestic violence support demand rapid response. Offices usually have too few intake specialists to begin with and too little time to process all comers. Yet the RCT's (and any other human subjects-based evaluation's) viability depends on a successful merger of consent processes and standard operating procedures. Successful legal aid RCT developers will take their time and think creatively about obtaining consent and accompanying documentation to satisfy IRB standards.

D. Responding to Stakeholder Resistance

Researchers willing to invest their time and energy in RCT program evaluations may spend months, even years, designing the study. Without notice, a lone stakeholder could then exercise veto power over the project. The researcher can do very little by this point to revive the idea. Consequently, I advise stakeholders to ask several rounds of questions in the early stages of RCT development. Lawyers instinctively recoil at resource allocation that departs from merit-based criteria or professional triage decision making. Representation decisions founded on anything other than clinical assessments is considered dereliction of duty. RCT ethics—namely the scarcity and equipoise conditions—should alleviate their deepest concerns. Lawyers considering rigorous evaluations are encouraged to think more broadly about service provision models in the process. On what basis can an attorney declare that selecting clients for representation based only on intake data is superior to a lottery? The answer returns us to the motivating question of this paper: are status quo procedures truly evidence-based in the scientific sense? When the answer is yes, the researcher should return to the office. There is no work to be done in the field. The answer is more often negative, and the legal profession should subject its core assumptions to more rigorous testing.

Field study developers usually encounter opposition to the very notion of randomization, because lawyers and other service providers believe that the study "deprives people of benefits" for which they otherwise might be eligible. A UNICEF-based evaluator more than accurately captures a common refrain:

> A typical discussion with those working on violence against children, poverty reduction, emergency response, nutrition, and more starts with colleagues telling me: "We want to rigorously test how well our programme works, but we don't want to do a randomized control trial (RCT)." For many in UNICEF, RCT is a bad word. *It conjures ideas of cold-hearted researchers arbitrarily withholding programme benefits from some households and villages for the sole purpose of racking up academic publications in journals no one will read*.[45]

It would be foolhardy to dismiss such an objection outright. People working on the frontlines have ample reason and every right to be suspicious of intrusive evaluations. Researchers advocating use of RCTs must take this skepticism seriously, and the best way to do so is to address it head-on. Doing so not only might increase the chance of conducting a legal RCT, it will also educate the profession on the foundational values of rigorous evaluation.

The key, but unproved, premise in the italicized portion of the UNICEF researcher's words is that the RCT involves "arbitrarily withholding programme benefits." Again, this sentiment is understandable. Interventions being evaluated are more likely than not appealing to some stakeholders along at least one dimension. (If not, it is hard to imagine why stakeholders would want to know whether it works.) Deeming the intervention a "benefit," though, presupposes the very conclusion the RCT places under the statistical microscope. It also violates the equipoise principle. Thus, the legal aid stakeholder is in a quandary. He wants to know whether some program positively impacts clients and simultaneously has declared it beneficial. Those two statements are obviously incompatible.

A second common retort is that people's lives are on the line in the legal aid world. Uncontested evictions or unanswered consumer debt claims destroy credit and dissolve families. And so, I would encourage researchers and stakeholders to recognize that high stakes in the law are reasons *to randomize and test*, not to avoid doing so. High stakes are why, for example, we insist on randomized evaluation of new cancer drugs; in such studies, the outcome variable studied is often patient survival for at least five years. When stakes are high, we should insist on rigorous evidence of effectiveness, not guesswork.

E. Willingness to Accept Unexpected or Undesirable Results

Finally, stakeholders committed to genuine program evaluations must steel themselves for unfavorable reviews, so to speak. The RCT could indicate no difference between a service provider's time-honored practice and a novel intervention. It could also suggest the intervention outperforms the previously untested approach. In both cases, legal aid stakeholders might experience significant buyer's remorse. Backward induction from unwelcome news to the RCT development stage therefore causes many attorneys to back away from an evaluation opportunity before the study launches. There is no causal relationship, however, between the RCT methodology and the likelihood of

[45] Tia Palermo, UNICEF, *Are Randomized Control Trials Bad for Children?*, https://blogs.unicef.org/evidence-for-action/are-random-control-trials-bad-for-children/ (emphasis added).

undesirable or unexpected findings. Identical qualitative outcomes are more than possible using less rigorous program evaluation tools. The only reason to worry about deriving the same results from an RCT is the reason to try an RCT in the first place: additional credibility from causal inference.

The point here is that bona fide curiosity is the *sine qua non* of rigorous evaluations. Legal aid stakeholders must be willing to absorb findings that upend traditional beliefs. After all, this process yields the very evidence-based practices lawyers claim they seek. Those with administrative control over evaluation decisions should not ignore the ancillary ramifications of the study's findings. Fear of budget cuts in particular exert strong emotional pulls to avoid a study, the results of which stakeholders cannot control. But the profession will never truly innovate if these political considerations alone preclude an RCT. Researchers should discuss openly and honestly the range of possibilities so that legal aid attorneys can make informed choices, too.

IV. CONCLUSION

Perhaps the most apt analogy for legal RCTs is the pharmaceutical drug trial. After initial exploration, researchers might surmise that a new treatment works reasonably well with minimal risk. They then can pilot the drug among a larger study population. After that, they proceed to the fully randomized trial. United States Food and Drug Administration approvals follow this protocol, and empirical researchers should strive to emulate that approach when considering a field RCT study in the law. It is instructive here to note that only 6% of drugs that professionals deem promising enough for multi-million-dollar investments in evaluation turn out, after RCT evaluations, to be effective for patients.[46] Professional judgment is therefore off the mark more than 90% of the time. We imagine similar numbers describe some legal phenomena, which increases the salience of an evidence base for common practices in the law.

As mentioned in the Introduction, this paper aims to do more than make the case for randomized experimentation in the law. Part II provides elementary, but clear, evidence for why the methodology is more likely to yield causal inference. The debate over how golden the RCT standard is will continue. In the meantime, there should be no question that the legal profession, and civil legal aid organizations more specifically, need more evaluation. Legal services policies and practices that were never subjected to "stress tests" or full scrutiny remain in force around the world. Continuing to deploy them without knowing whether and to what extent they actually improve clients' lives is not rational. The best course correction, and the agenda that this paper represents, is to elevate rigorous evaluation in legal services to the vaunted position it occupies in the medical and pharmaceutical industries.

I encourage service providers to use the lessons herein and to consider both reviewing current procedures and assessing whether they can try something new. With

[46] *See* Chi Heem Wong, Kien Wei Siah & Andrew W. Lo, *Estimation of Clinical Trial Success Rates and Related Parameters*, 20 BIOSTATISTICS 273, 280 (2019).

respect to the latter, organizations ideally would embed an evaluation *at the same time as it pilots the innovation.* As soon as new paradigms take hold, become familiar, and convince users (through sense impressions) that they are effective, discarding them can be near impossible. As a result, service providers will preferably start building an evidence basis from the moment an idea is introduced. For lawyers, this means early collaboration with empirical researchers on an evaluation plan coinciding with the program's unveiling. For government entities, it might entail statutory provisions that tie funding for new projects to real-time piloting and testing. Evidence-based practices have the potential to revolutionize the civil access to justice movement. Legal professionals are now tasked with seeing that agenda through and embracing evaluation methods that make evidence-based practices the new standard.