2008

Randomized Legal Experimentation

Michael B. Abramowicz
George Washington University Law School, abramowicz@law.gwu.edu

Randomized Legal Experimentation

Michael Abramowicz
George Washington University

Social scientists have performed and analyzed a number of randomized studies of policies, but the legal literature has not addressed whether and how the legal system should incorporate experimental methods. This Article identifies several benefits of randomized legal experimentation and argues that these benefits supports self-executing experiments, whose results would lead to policy changes agreed upon in advance. Randomized experiments can generate information, and self-execution can help ensure that this information affects the policy process. Such experiments may be easier to enact than other legal reforms, because each side of a policy debate may believe that an experiment is likely to support its cause. This Article argues that administrative law doctrine should be more deferential to agency decisions to perform such experiments than to enact policies without evidence from randomized experiments. The Article describes the advantages and limitations of randomization, and explores ethical and equality-based arguments against experimentation.

I. INTRODUCTION

Policymakers and commentators frequently refer loosely to new laws and legal institutions as "experiments,"[1] but in contrast to medical experimentation,[2] these innovations almost never randomly designate treatment and control groups. There have been a handful of exceptions since 1968,[3] randomized "social experiments" on the impact of government policies. But the legal literature has virtually ignored them. Legal scholars have discussed the results of particular social experiments,[4] and occasionally have commented that additional social experiments could provide useful information in one field or another.[5] But they have not addressed the normative question of whether the legal system should generally seek to incorporate experimental methods, and if so, what approaches the legal system might take to maximize the chance that experiments will improve policy.

Perhaps viewing past social experiments as only incidentally involving government as the sponsor of research, legal scholars have been content to leave analysis of them to the social science literature. Unsurprisingly, that literature, like the economics literature on "field experiments,"[6] has focused on what social scientists have learned about behavior from previous

---

[1] *See, e.g.*, Orit Fischman Afori, *Reconceptualizing Property in Designs*, 25 CARDOZO ARTS & ENT. L.J. 1105, 1151 (2008) (referring to a statute providing intellectual property protection for vessel hulls as a "legal experiment"); Theodor Meron, *Reflections on the Prosecution of War Crimes by International Tribunals*, 100 AM. J. INT'L L. 551, 551 (2006) (referring to the Nurenberg and Tokyo war crimes tribunals as "a bold legal experiment"); Alan Milner, *Restatement: The Failure of a Legal Experiment*, 20 U. PITT. L. REV. 795 (1959) (characterizing restatements of law as a failed experiment). The most prominent academic account of experimental approaches to government also defines experimentation broadly, mentioning randomization as a possible ingredient of experimentation only once. *See* Michael C. Dorf & Charles F. Sabel, *A Constitution of Democratic Experimentalism*, 98 COLUM. L. REV. 267, 348 (1998) (noting that systems for evaluating experiments "can themselves be benchmarked, and . . . can be combined with random-assignment experiments and other familiar methods of evaluation").

[2] For a historical discussion of the introduction of randomization into statistical analysis in medicine, see Tar Timothy Chan, *History of Statistical Thinking in Medicine*, *in* ADVANCED MEDICAL STATISTICS 3, 11-14 (Ying Lu & Ji-Qian Fang eds., 2003). *See also* Ronald A. Fisher, *The Arrangement of Field Experiments*, 33 J. MIN. AGRIC. GREAT BRITAIN 503 (1926) (introducing the idea of the random trial).

[3] A doctoral student, Heather Ross, developed the idea for the experiment, on the effect of a negative income tax, and then received governmental funding. The experimental results are reported in three volumes. *See* 1-3 THE NEW JERSEY INCOME MAINTENANCE EXPERIMENT (Harold W. Watts et al., 1976-1977). Useful summaries of the experiment are in SOCIAL EXPERIMENTATION 95-143 (Jerry A. Hausman & David A. Wise eds., 1985); and DAVID GREENBERG ET AL., SOCIAL EXPERIMENTATION AND PUBLIC POLICYMAKING 111-64 (2003).

[4] *See, e.g.*, Machaela M. Hoctor, Comment, *Domestic Violence as a Crime Against the State: The Need for Mandatory Arrest in California*, 85 CAL. L. REV. 643, 655-57 (1997) (commenting on a Minneapolis experiment with randomized mandatory arrest of alleged domestic violence perpetrators).

[5] *See, e.g.*, Laurens Walker, *Perfecting Federal Civil Rules: A Proposal for Restricted Field Experiments*, LAW & CONTEMP. PROBS., Summer 1988, at 67 (proposing randomized experiments on procedural rules).

[6] For an overview of the recent research, see Glenn W. Harrison & John A. List, *Field Experiments*, 42 J. ECON. LIT. 1009 (2004).

experiments and how to design future experiments to maximize such information.[7] Scholars have paid some attention to the factors that influence governments' decisions whether to create social experiments[8] and ultimately to change policy as a result of them,[9] as well as to the ethical considerations inherent in social experiments.[10] They have not, however, theorized about the role that randomized experiments might and should play generally in governmental processes and legal doctrine.

Perhaps in part as a result of this scholarly neglect, past social experiments have clustered in specific policy areas. As the label "social experimentation" suggests, most of the experiments have been in the area of social services, testing whether expenditures on entitlements succeed in achieving social goals such as reducing poverty.[11] For example, a current experiment, executed under a Medicare statute requiring randomized testing of programs,[12] is assessing whether telephone contact by nurses to at-risk Medicare patients will reduce program costs.[13] Another class of randomized studies test criminal justice policies.[14] A rare exception outside these two areas was a set of experiments on electricity pricing.[15] None of the experiments randomly vary legal rights and obligations of ordinary citizens and entities in areas such as taxation or patent

---

These experiments test propositions such as whether offering matching donations increases charitable contributions, *see, e.g.*, Stephan Meier, *Do Subsidies Increase Charitable Giving in the Long Run? Matching Donations in a Field Experiment*, 5 J. EUR. ECON. ASS'N 1203 (2007) (finding that offering matching donations to potential donors increases donation in the short run but decreases donation in the long run), or whether increased teacher attendance improves students' educational performance, *see* Esther Duflo & Rema Hanna, *Monitoring Works: Getting Teachers to Come to School* (NBER Working Paper No. 11880, Dec. 2005) (finding that it does).

[7] These questions are, for example, the focus of SOCIAL EXPERIMENTATION, *supra* note 3.

[8] *See* GREENBERG ET AL., *supra* note 3, at 46-48.

[9] *See* GREENBERG ET AL., *supra* note 3, at 48-63 (assessing the factors affecting the influence of experiments in the policymaking process). This analysis, however, does not consider experimental approaches developed in this Article, such as the possibility of self-executing experiments.

[10] *See* ETHICAL AND LEGAL ISSUES OF SOCIAL EXPERIMENTATION (Alice M. Rivlin & P. Michael Tipane eds., 1975).

[11] "[M]ost social experiment test programs are targeted at persons or families who are somehow disadvantaged, particularly in terms of having low incomes." GREENBERG ET AL., *supra* note 3, at 26.

[12] Medicare Prescription Drug, Improvement, and Modernization Act of 2003, Pub. L. No. 108-173, § 721 (codified at 42 U.S.C. § 1395b-8) (requiring "development, testing, and evaluation of chronic care improvement programs using randomized controlled trials").

[13] *See* NANCY MCCALL ET AL., CENTERS FOR MEDICARE & MEDICAID SERVICES, EVALUATION OF PHASE 1 OF MEDICARE HEALTH SUPPORT (FORMERLY VOLUNTARY CHRONIC CARE IMPROVEMENT) PILOT PROGRAM UNDER TRADITIONAL FEE-FOR-SERVICE MEDICARE (June 2007); Reed Abelson, *Medicare Finds How Hard It Is To Save Money*, N.Y. TIMES, Apr. 7, 2008, at A1 (describing the program).

[14] *See generally* David F. Farrington & Brandon C. Welsh, *A Half Century of Randomized Experiments on Crime and Justice*, 34 CRIMINAL JUSTICE 55 (2006) (providing an overview of randomized criminal justice experiments, the first of which was initiated in 1951 and reported on in 1978).

[15] *See* SOCIAL EXPERIMENTATION, *supra* note 3, at 11-53; *see also* RESEARCH TRIANGLE INST., ANALYTICAL MASTER PLAN FOR THE ANALYSIS OF THE DATA FROM THE ELECTRIC UTILITY RATE DEMONSTRATION PROJECTS (1978).

law. Instead, they focus on possible provision of new services or on those who might be thought of as forfeiting rights by committing crimes.

This Article advances a case for "legal experimentation," using this phrase rather than "social experimentation" to highlight that such experimentation might be considered in almost any legal context, from civil procedure to tort law, from employment law to securities law. Experiments have the potential not merely to be governmentally funded academic exercises, but to serve as integral components of the legal process. The principal traditional justification for legal experimentation, that experimentation creates information, both overstates and understates its potential. Experimental data will rarely give unambiguous answers to multi-dimensional policy questions. The results of random experimentation will become just additional pieces of information available to decisionmakers, and there is little reason to expect that the influence of information will be proportional to its quality.[16] But experiments could have greater impact if governments were to make policy conditional on experimental results. This Article will argue for *self-executing* randomized legal experiments. A self-executing experiment either could specify ex ante the policy effects of particular results, or, as in the Medicare experiment, could require independent decisionmakers to make policy changes based on the experiment.[17] The hope is to nudge policy at least a small distance in what will generally be the right direction while avoiding some of the public choice hurdles and legislative inertia that often frustrate change.

A self-executing experiment, of course, would still require legislative or administrative authorization, and so it cannot avoid these obstacles altogether. But if through gradual steps randomized self-executing experiments become sufficiently familiar that they no longer seem strange,[18] then a culture of random legal experimentation might slowly emerge. Legal experiments should be easier to enact in this culture than are legal reforms in our present legal culture. A marginal decisionmaker, uncertain whether to support a program, should be more willing to favor it when the program will continue if and only if it turns out to be successful. Supporters of a program, meanwhile, may find it difficult to oppose a measure that would condition continuation of the program on confirmation of its success. Even opponents of moving

[16] *See infra* Part II.A.1.

[17] *See infra* notes 66–68 and accompanying text (describing self-executing features of the Medicare experiment).

[18] *See infra* Part IV.B.2.

the law in the direction of an experiment might nonetheless be willing to support the experiment if they believe that it will turn out to be a failure.

These effects may be sufficient to promote experimentation on the margins even in today's legal culture, but in a mature legal experimental culture, norms could emerge that could further facilitate experimentation and legal change. We can imagine, for example, bidirectional self-executing experiments, which would move the law automatically in the direction of the experiment if it proves successful, and in the opposite direction if it fails. If ideological opponents genuinely disagree about the effects of potential policies, such experiments can seem beneficial ex ante to all, increasing the gains from political trade. Such experiments channel ideological disagreement, increasing the possibility of legal change rather than hardening impasse. Such experiments seem unlikely in our current legal system, but growing comfort with experimentation, randomization, and self-execution might someday make it so that opposition to such an experiment would be perceived as indicating that a policymaker lacked confidence in empirical claims advanced on behalf of a proposal.

As a theoretical matter, this logic may best help justify creation of self-executing experiments in a legislature. But as a practical matter, self-executing legal experiments in the foreseeable future are more likely to originate in administrative agencies, though sometimes, as in the Medicare case, as a result of a legislative requirement. This Article is the first to consider how administrative law doctrine should contemplate legal experiments, and it argues that such doctrine should expressly recognize the benefits of experimentation. Courts would provide experimenting agencies more latitude when reviewing agency procedures, as well as factual, legal, and policy determinations, especially for programs that are randomized and self-executing, and even more so when self-execution is bidirectional. When agencies undertake legal experiments, there is generally a reduced risk that they are pursuing a particular ideological result, and the objectivity of randomized analyses should similarly reduce the potential for ideology in judicial review of agency decisions.

These arguments highlight that we cannot consider legal experiments solely as a social scientist might. Rather, we must consider legal experimentation as a mechanism of the policymaking process, an imperfect device for converting scientific knowledge into law. Sometimes, the criteria of scientific usefulness and legal practicality point in different directions.

An experiment might be beneficial even if its results add little to social science knowledge; a simple randomization scheme may be beneficial even where econometricians would prefer a more elaborate treatment design; and an experiment might compare two legal approaches varying along a number of dimensions even though this may make the results difficult to interpret. The trade-off between scientific usefulness and legal practicality need not always be resolved in favor of a policy-oriented approach, however, and some legal problems, even where the fundamental policy questions are empirical, do not lend themselves to randomized experimentation. This Article will draw on a wide variety of hypothetical experiments, but not all of these hypotheticals are designed to illustrate virtues of this approach. Many are chosen because they illustrate challenges for legal experimentation.

The Article proceeds as follows. Part II assesses the benefits of legal experimentation generally, arguing that it can improve information relevant to the policymaking process (though this does not guarantee that the information will be used), that it can increase the chance of moving the law in a direction that a median legislator would favor, and that it increases electoral accountability by making legislators' goals more transparent. It also assesses the prospects for legal experimentation by legislatures, administrative agencies, and courts, and it compares legal experimentation to alternative means of securing more independent and scientific decisionmaking, such as sunset clauses, reporting requirements, delegation to administrative agencies, and privatization. Part III explains from both statistical and policy perspectives why randomized experiments are likely to be more useful than nonrandomized evaluations, while also considering a variety of hazards of randomized experiments and what these hazards mean for legal experiments. Finally, Part IV provides some guidelines for legal experimentation and directly considers ethical and other objections to randomization and to self-execution.

II. THE POLITICAL ECONOMY OF LEGAL EXPERIMENTATION

This Part identifies benefits of experimental government in various institutional contexts and compares experimentation to other devices serving similar goals. Part II.A explains that in a pluralist policymaking environment, experimentation can improve political debate and enhance policy, particularly if randomized, self-executing experiments became entrenched in political practice. Part II.B assesses the suitability of legislatures, administrative agencies, and courts for

instituting beneficial experiments, and Part II.C compares experimentation to sunset clauses, reporting requirements, delegation to independent agencies, and privatization.

A. The Benefits of Experimental Government

Experimental government can improve policy in several ways: experiments can generate information, experiments can facilitate the passage of reforms, and debate about experimentation can make legislators' goals more transparent. The challenge is not merely producing information, but also beneficially influencing the policy process. Randomization can help simplify information and make the purpose of an experiment, and thus legislators' aims, more apparent. Self-execution can guarantee that experimentation influences the public, turning disagreement into a force that makes policy change easier, while also helping to make legislators' goals explicit.

1. Information Generation

The most obvious defense of legal experimentation is that it can produce information not otherwise obtainable. This insight applies even to nonrandom experimentation. Implementation of a proposed policy will usually provide better evidence of the policy's effects than speculation. Consider, for example, no fault automobile insurance. When no jurisdiction had enacted no fault automobile insurance, academics could speculate about its likely effects, such as an increase in accidents and a per case decrease in litigation expenses.[19] But attempts to gauge these effects' magnitudes involve guesswork.[20] Once a jurisdiction adopts no fault insurance, observers can casually assess the empirical effects. Nonetheless, such assessments are prone to error. For example, a rise in accidents after adoption of no fault insurance could be a result of some independent cause, could be related in an indirect way (for example, if a new political party assuming office and had enacted numerous policies including no-fault), or a transient statistical fluke. This highlights limitations of a nonrandom experiment: the treatment cannot be isolated and the decision to treat may be an endogenous function of the policy environment.[21]

---

[19] *See, e.g.*, Guido Calabresi, *The Decision for Accidents: An Approach to Nonfault Allocation of Costs*, 78 HARV. L. REV. 713 (1965) (assessing the advantages and disadvantages of a no-fault liability system).

[20] *See* Richard A. Posner, *Guido Calabresi's* The Costs of Accidents*: A Reassessment*, 64 MD. L. REV. 12, 18-19, 21 (2005) (noting that Calabresi's work makes numerous theoretical claims without empirical support).

[21] Careful researchers attempt to account for these limitations. *See, e.g.*, J. David Cummins et al., *The Incentive Effects of No-Fault Automobile Insurance*, 44 J.L. & ECON. 427 (2001) (offering an empirical analysis of no-fault insurance that takes into

Suppose, in contrast, that a state randomly selected half of its citizens to switch to a no fault insurance regime, while the other half continued to have fault-based insurance,[22] and allowed insurers to charge differential prices. The state might then monitor variables such as insurance premiums, accident rates, and litigation costs, for example by imposing a reporting requirement on insurers. There might be difficulties in extrapolating what these variables would be if the jurisdiction switched entirely to no fault insurance,[23] but this approach helps to overcome the problems of an uncontrolled experiment. Any other simultaneous legal or nonlegal changes in the state cannot explain any significant differences between the two populations. When properly randomized, the treatment decision is exogenous. Even if there is only a small chance that such an experiment would succeed, this possibility could trigger a lasting legal change, so the additional information gained could produce expected social benefits outweighing expected costs.

Commentators have occasionally remarked that social experiments will benefit society only if they influence policy.[24] David Mundel, for example, has argued that to assess experiments we must ask whether they answer important policy questions, whether the answers are understandable, and whether understandable answers can alter the beliefs of policymakers.[25] Even this breakdown is incomplete, because policymakers with changed beliefs might not act on those beliefs, for example because interest groups oppose a policy change. The focus of the social experimentation literature has been how to design experiments to produce reliable policy answers,[26] perhaps on the implicit premise that better studies will be more likely to clear each of

---

account the endogeneity of the decision to adopt such insurance). Statisticians might debate, however, whether any such attempt is successful.

[22] A complication with this approach is that individuals with fault-based insurance might end up in accidents with individuals with no fault insurance. The legal regime must adopt some approach to dealing with such accidents. For example, the law might pick one approach or the other to use in such cases, or it might stipulate that only a party with fault-based insurance can be liable for the other party's injuries on account of fault. In a federal system, the law must deal with such incompatibilities anyway. *See, e.g.*, Michael W. Mengis, *Conflict of Laws: Insurance*, 47 LA. L. REV. 1213, 1220-23 (1987) (detailing conflict-of-law rules involving automobile accidents). With a large enough treatment group, it should still be possible at least to obtain an approximate sense of the effects of a shift from fault to no fault insurance.

[23] During an experiment, any given driver may be less likely to know what regime applies to that driver than if a legal change were permanent.

[24] *See* Jerry A. Hausman & David A. Wise, *Introduction*, *in* SOCIAL EXPERIMENTATION, *supra* note 3, at 1, 2 ("[S]ince, to a large extent, the policy questions the experiments were designed to answer still have not been decided, the final accounting of the worth of the experiments in helping to decide the course of public policy is probably fairly far off into the future.").

[25] David S. Mundel, *The Use of Information in the Policy Process: Are Social-Policy Experiments Worthwhile*, *in* SOCIAL EXPERIMENTATION, *supra* note 3, at 251, 252.

[26] Hausman & Wise, *supra* note 24, at 1 ("[T]he most important question is whether the experiments have been successful in their primary goal of providing precise estimates of the effects of a proposed government policy.").

these hurdles. Some relaxation of the goal of scientific purity, however, could sometimes facilitate each of Mundel's steps and promote the ultimate goal of creating new policy.

First, experiments, though reducing uncertainty, will rarely provide definitive answers about whether particular legislation should be adopted. The treatments employed in a random experiment will often not correspond unambiguously to policy options. For example, the RAND Health Insurance Experiment assessed coinsurance requirements' effect on health services utilization and outcomes.[27] But no legislation focused specifically on coinsurance, and those who opposed the experiment pointed out that "the insurance plans included in the HIE did not correspond directly to the policy options being considered by Congress at the time."[28] The experiment may have influenced individual health care reformers.[29] But it is doubtful that it materially affected the viability of any legislation. A social experiment focusing on one dimension of a problem leaves room for debate about the effects of a policy initiative with many other dimensions. This suggests one difference between scientific and legal goals in experimental design. The RAND experiment targeted concrete, scientifically manageable questions.[30] An experiment designed around a legal proposal—for example, randomizing some people to a single payer plan and requiring those individuals to pay any higher taxes needed to fund it—might have been less scientifically useful, because it would have been difficult to identify which elements of the plan produced certain results. But it might have been more legally useful, providing direct assessments of a proposed legal reform taken as a whole.[31]

---

[27] *See* Joseph P. Newhouse et al., *Some Interim Results from a Controlled Trial of Cost Sharing in Health Insurance*, 305 NEW ENG. J. MED. 1501 (1981) (providing the first reported results on the experiment).

[28] GREENBERG ET AL., *supra* note 3, at 73.

[29] For example, economists helping to develop the details of President Clinton's health plan used the RAND results in setting parameters in simulation models designed to assess the ramifications of different options. *Id.* at 77-81. It seems plausible that if that health plan had been enacted into law, some of the law's features might have been traceable to the results of the RAND experiment.

[30] At least some commentators have drawn sweeping conclusions from the RAND Experiment. For example, a major finding of the experiment was that greater use of health services did not improve health outcomes. Some commentators have argued that this suggests that there might be overuse of health care and that the government should not be subsidizing it. *See, e.g.*, Robin Hanson, *RAND Health Insurance Experiment*, http://www.overcomingbias.com/2007/05/rand_health_ins.html,

[31] In other experiments as well, critics have argued that the experiments' focus on a few variables has made them useless for broader policy analysis. Harvey Rosen, for example, argued that results from a randomized housing experiment "offer no particular advantages" over nonrandomized observational data. "[H]ousing behavior is so complex and the policy environment so uncertain that simple comparisons of experimental and control groups are unlikely to be of much interest," he argued. Harvey S. Rosen, *Housing Behavior and the Experimental Housing-Allowance Program: What Have We Learned*, *in* SOCIAL EXPERIMENTATION, *supra* note 3, at 55, 72. "Rather, the data must be interpreted with the help of theoretical and statistical models." *Id.*

Second, scientific and legal goals may also be in tension for Mundel's next criterion, whether the results are understandable. An advantage of random experimentation is that it can make statistical results accessible.[32] Even a statistical naïf can informally compare treatment and comparison groups. Sophisticated econometric techniques can correct for potential problems in random experimentation.[33] The typical scientific instinct would be to apply such corrections, but these techniques are not methodologically simple or uncontroversial. Similarly, statisticians might test subsamples to determine whether the intervention has different effects on different individuals or communities.[34] What makes results more understandable and meaningful to a neutral expert statistician may make the results less useful in the policymaking process.[35] Sometimes, policy should depend on analysis beyond a raw comparison of treatment and comparison groups. But given the difficulties of identifying the objectively best way to massage data, it often may be preferable for a legal system to insist in advance on a single relatively good comparison rather than to strive for a perfect one.[36]

Third, even if results are understandable to a public official who takes the time to study them, they might not affect public officials' beliefs. Henry Aaron has noted skeptically, "In the policy forum, social science research is 'evidence' to be introduced as part of a frequently multifaceted adversary process in which each side builds its case to make it as persuasive as

---

[32] Mundel notes, "If properly conceived and analyzed, social experiments can result in simply stated and easily understood conclusions." Mundel, *supra* note 25, at 254. The problem is that if there are multiple competing analyses, conclusions can become muddier.

[33] An example of such a problem is the possibility of differential attrition in the treatment and comparison groups. *See infra* Part III.B.1.b.i.

[34] A problem with this approach is that with enough subsamples tested, there is a high probability that a simple regression analysis will produce spurious statistically significant results on some subsamples.

[35] Hausman and Wise recognize this point: "[T]he use of complex structural models to analyze the data from social experiments … are often in contradiction to the primary motivation for the experiments and thus subvert their intent; they are often inconsistent with the raison d'être of experiments." Jerry A. Hausman & David A. Wise, *Technical Problems in Social Experimentation: Cost Versus Ease of Analysis*, *in* SOCIAL EXPERIMENTATION, *supra* note 3, at 187, 190. The primary motivation they allude to is the goal of producing simple and clear answers to scientific problems.

[36] Some observers have complained that the data from past experiments were less useful as a result of attempts to make the data useful for the policy process. Paul Joskow, for example, assessed experiments on the effects of charging higher electricity prices at times of peak demand. *See* Paul L. Joskow, *Comment*, *in* SOCIAL EXPERIMENTATION, *supra* note 3, at 42. The policy debate was between environmentalists, who insisted that higher prices could reduce peak demand "so as to reduce the need for additional power plants," *id.* at 44, and electric utilities that were skeptical that demand would be responsive to prices. Joskow complains that the experiments thus ended up focusing on what should have been uncontroversial from an economic perspective, that "the elasticity of demand for electricity was negative." *Id.* Yet one might argue that from a policy perspective, even if disagreement focused on an elementary question of economics, it was useful for the experiments to target this question directly, rather than to introduce many experimental variations that would produce more information for interested scientists yet less understandable results for public policy.

possible."[37] Mundel more charitably describes policy makers as "classically Bayesian—entering a problem with an estimate of the likely outcome and an estimate of the variance of uncertainty surrounding the likely outcome," and then changing these estimates based on evidence emerging from experiments.[38] But the degree to which policymakers adjust their beliefs may not be proportionate to the quality of evidence,[39] especially if policymakers are not perfect Bayesians, but instead, as the cognitive psychology literature suggests,[40] filter new information to fit with their pre-existing views. Cognitive psychology also suggests that more salient experiments may have greater influence on policymakers' views.[41] Salience will be maximized with a prior agreement on the conditions for determining whether an experiment will be counted as a "success" or a "failure." There is some danger that simplistic labels can lead policy in the wrong direction, but this danger must be balanced against the chance that subtle findings will have no policy influence.

Fourth, self-executing experiments substitute for the final step that Mundel's analysis neglected: the conversion of an experiment into legal policy. Even a policymaker who is genuinely swayed by a non-self-executing experiment might not admit it,[42] for fear of seeming inconsistent or of alienating supporters. A policymaker might point to the results of other studies, even less rigorous ones.[43] Or, a policymaker might note independent justifications for the

---

[37] Henry Aaron, *Comment*, *in* SOCIAL EXPERIMENTATION, *supra* note 3, at 272, 275.

[38] Mundel, *supra* note 25, at 255.

[39] Policymakers may also pretend not to adjust their views. Lant Pritchett has suggested that perhaps one reason for experiments' lack of influence is that policymakers may not support rigorous analysis of programs they support because they fear that doing so will make it more difficult to these programs. *See* Lant Pritchett, *It Pays to be Ignorant: A Simple Political Economy of Rigorous Program Evaluation*, 5 J. POL'Y REFORM 251 (2002); *see also* Michael Kremer, *Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons*, 93 AM. ECON. REV. 102, 105 (2003) (discussing Pritchett's view).

[40] *See, e.g.*, Donald C. Langevoort, *Behavioral Theories of Judgment and Decision Making in Legal Scholarship: A Literature Review*, 51 VAND. L. REV. 1499, 1506 (1998) (discussing cognitive dissonance).

[41] *See* Russell B. Korobkin & Thomas S. Ulen, *Law and Behavioral Science: Removing the Rationality Assumption from Law and Economics*, 88 CAL. L. REV. 1051, 1090 (2000) ("[A]n understanding of the availability heuristic can provide a caution to policymakers tempted to enact new regulatory regimes in response to highly available information concerning health or safety risks.").

[42] Some policymakers may offer candid acknowledgments that their initial instincts were wrong. Sen. Daniel Patrick Moynihan, for example, who championed experiments on a negative income tax, admitted that the results seemed to cut against his initial views. *See* GREENBERG ET AL., *supra* note 3, at 144 (quoting Sen. Moynihan as saying, "[W]e were wrong about a guaranteed income! Seemingly it is calamitous…. Such is not the state of the science, and it seems to me we are honor-bound to abide by it for the moment."). But Sen. Moynihan had achieved renown as a scholar, and might have been in a better position to admit being persuaded by new data than other politicians.

[43] Flawed studies may receive a great deal of political and media attention. *See, e.g.*, Richard Lempert, *Strategies of Research Design in the Legal Impact Study: The Control of Plausible Rival Hypotheses*, 1 LAW & SOC'Y REV. 111, 125 (1966-1967) (noting that studies assessing the effect of a single law in a particular jurisdiction "tend[] to be very misleading to the public at large, yet … often are publicized widely").

policymaker's initial preferences. Self-execution overcomes disingenuous resistance to an experiment's conclusions, but it may appear to make the final policy decision less scientific. A self-executing experiment determines policy based on a mere proxy for policy effectiveness. It only changes the policy baseline, but an official label of an experiment as being a success or a failure may obscure more subtle information revealed by an experiment. On the other hand, policymakers often pay more attention to whether particular variables are significant than to the magnitudes of the corresponding statistical effects,[44] when they pay attention to experimental results at all.[45] Whether a self-executing experiment produces a numeric result exceeding a predetermined threshold may be no less effective a proxy.

The legal system need not always rely on simplified self-executing experiments for information to have an effect on public policy. Sometimes, policymakers might incorporate experimental results in a thoughtful, nonideological way. Perhaps this is true of the current Medicare experiments, although the creation of an institutionalized experiment generation and evaluation mechanism may itself be conceived as akin to a self-executing experiment. At other times, any attempt to reduce a particular experiment ex ante to a single metric labeled "success" or "failure" may itself be so flawed that there is an unacceptably high risk of incorporating bad information into the policy process. But often for experiments producing information in political settings, less may be more.

2. Reform Facilitation

The analysis so far has focused on an experiment's aftermath, rather than on its initiation. This section addresses whether experimentation promotes or substitutes for legal reform. Maybe "experiments are conducted to replace real action with symbolic action and thereby serve as an inherently conservative form."[46] Sometimes an experiment might be used to take an issue off the public policy agenda until political storms pass.[47] But this could backfire, because experiments

---

[44] GREENBERG ET AL., *supra* note 3, at 144 (reporting that in the income maintenance experiments, "evidence on the existence of the effect tended to receive much more notice within political circles than evidence on the size of the effect," even though the intent of many researchers was to show that while some effect existed, it was relatively small).

[45] Some apparently important results in social experiments have received little attention. For example, one result of the income maintenance experiments was that subsidized training voucher programs, giving recipients choice of what training to undertake, failed to increase earnings. *See id.* at 129.

[46] *Id.* at 47.

[47] In experiments on whether welfare recipients could be induced to take jobs, "welfare directors in states that were testing welfare-to-work innovations on a pilot basis liked being able to hold off state legislatures until they could learn about the

provide "rigorous tests of policies" and thus may further support for reform.[48] This is especially so with random experimentation, which produces relatively rigorous and comprehensible results. Use of experiments to defuse calls for reform might be unproblematic anyway. First, delaying decisions after a crisis may ensure that policy is based on more rational reactions, relatively free from cognitive biases overemphasizing factors of high salience.[49] Second, other mechanisms already allow policymakers to put off issues. Policymakers can form study committees, or pass symbolic reform. Experiments probably do not much increase the opportunities that policymakers have to avoid contentious issues.

More importantly, there are more powerful forces by which experimentation could encourage legal reform. Consider a marginal policymaker, who is unsure about whether to support a particular policy change. An experiment is an initiative that this policymaker might well support, subject to concerns such as cost, because of the information that it produces. Meanwhile, policymakers who favor some legal reform may support an experiment if they are not sufficiently numerous to push through the legal change on their own and if the experiment may sway enough votes. On this theory, policymakers are likely to support precisely those experiments that have the best chance of producing information that ultimately will lead to reform.

Where there are competing proposals on an issue, experimentation may receive even more support. Suppose that liberals favor Plan L, conservatives favor Plan C, and moderates are uncertain, and suppose further that in the absence of experimentation, the status quo will be maintained. Then, all three camps might well favor experimentation on some disputed empirical aspect of the issue, because each side might anticipate that the experiment will verify its empirical claim and bring the uncommitted to that side. With experimentation, disagreement on empirics counterintuitively may make political bargains easier to strike. It may then seem surprising that there is so little experimentation, particularly so little of the random experimentation that seems especially likely to influence the uncommitted. A partial explanation is that this theory works only if the experiment has some realistic chance of influencing enough

---

effectiveness of the innovation." *Id.* at 219.

[48] *Id.*

[49] *See generally* Timur Kuran & Cass R. Sunstein, *Availability Cascades and Risk Regulation*, 51 STAN. L. REV. 683 (1999) (describing how the availability heuristic can lead to impetuous regulation).

moderate policymakers to lead to permanent enactment. The typical experiment may be of such low political salience that this will rarely be true.

Indeed, non-self-executing randomized experiments may be of lower salience than nonrandomized experiments that change the status quo for an entire population. Conceptualization of experiments as academic exercises may partially explain why they are generally no larger than necessary to produce needed information for policymakers. Random experimentation may appear from a budgetary perspective to crowd out other academic research. Yet many policies, even with uncertainty about empirical effects, are enacted either on a nonrandomized or entirely nonexperimental basis, with far larger budgetary consequence. If randomized experiments seemed a natural part of policy development, then experimental proposals might be evaluated using ordinary informal cost-benefit criteria, rather than competing for scarce research funding. At sufficient scale, the results of randomized experiments might then penetrate general political discourse. But randomized experiments are unlikely to seem natural until they are used more often, a chicken-and-egg problem.

The possibility of self-executing experiments could provide an escape from this dilemma. Ideological opponents might agree to a self-executing experiment even where in the absence of self-execution, the result of an experiment would not seem likely to affect policymakers' decisions. Self-execution will be attractive when ideological opponents have genuinely different beliefs about expected empirical outcomes. With self-execution, each side seeks not necessarily to persuade the other, but to take advantage of the other side's flawed expectations. Admittedly, self-executing experiments face a chicken-and-egg problem of their own. The self-executing nature of an experiment may itself provoke resistance.[50] A principal goal of this Article is to build the case for self-executing experiments, but no academic article seems likely to change convention enough on its own. If self-executing experiments are to become a prominent part of the policy landscape, they will likely emerge in small steps, such as the ongoing Medicare experiment.[51] Such experiments, even if not scientifically definitive, might come to be perceived as generally shifting policy baselines in a beneficial direction.

[50] *See infra* Part IV.B.2.

[51] *See supra* notes 12–13 and accompanying text; *infra* notes 64–68 and accompanying text.

The changes that self-executing experiments effect will not always be beneficial. A plausible assumption, however, is that when policymakers are able to approve a policy through standard constitutionally specified means, welfare is more likely to increase than decrease. On this assumption, if experimentation ultimately facilitates permanent legal changes, we tally those changes as expected improvements. This assumption, by refusing to privilege the status quo over change, implicitly rejects the possibility that a "law of unintended consequences" means that well-intentioned experiments will have harmful results more often than not. If legal change is generally undesirable, then any mechanism for increasing consensus for legal change will be harmful. Arguably, self-execution increases the dangers of unintended consequences, because policy is specified on the basis of experiments that have not yet occurred and cannot yet be fully evaluated. Randomization, however, at least alleviates the dangers through rigorous testing.

3. Goal Transparency

Experimentation also may promote public accountability by making policymakers' goals more transparent to constituents. A policymaker may announce concerns with some particular public policy variable—say, whether reduction of welfare will promote job creation—but agreeing to an experiment that will test the effect of a policy makes this alleged commitment more credible. As importantly, a policymaker who claims to oppose a policy because of a concern about a potential effect may find it hard to oppose an experiment that would reveal whether this effect exists. Some policymakers might argue that they think an idea is so bad that even experimenting with it is ill-advised, but especially if experimentation became more common, opposition might signal that a policymaker's true concern is not the one publicly stated.

Randomized experiments can augment goal transparency, because randomization enhances the transparency of the experiments themselves. Randomized experiments can allow for simple comparisons of the treatment and control populations. A proposal to conduct a randomized experiment can thus make clear what is being tested. When a policymaker announces a proposal for a randomized experiment of a policy that the policymaker claims will have a particular effect, the public can have relatively high confidence that the policymaker genuinely believes this. Because randomization constrains the degree to which a policymaker

can spin experimental results,[52] a policymaker who supports a new permanent policy will be willing to accept randomized experimentation only if there is a sufficiently high chance that the experiment will be successful. The willingness of a policymaker who supports a program to subject it to randomized testing supports the honesty of the policymaker's claims. Opponents of a particular initiative similarly might suggest randomization to enhance their empirical claims.

Proposals for random experimentation thus can be useful even if the results of experimentation could be readily predicted by policy insiders. Such proposals focus debate around the empirical effects of a policy and allow the public to gauge better what insiders think the effects of a policy might be. Admittedly, if randomized experiments are viewed as mere academic studies, a policymaker's proposal of a randomized experiment might receive little attention in a broader policy debate. Meanwhile, a proponent of a policy might not see any political need to respond to an opponent's suggestion of an experiment. The goal transparency advantages of experimentation may thus require experimentation to become more deeply embedded in policymaking culture.

If such a culture did develop in some agency or legislature, then debate might ensue not just about whether there should be an experiment, but also about what precisely should be tested in the experiment. The possibility of experimentation should lead both sides to trim extravagant policy claims. If either side endorses only a partial step to a controversial policy, the other side might point out that this suggests fear about a test of a fully-implemented policy. Proponents thus will not want to advocate a policy that opponents would be all too happy to test in a high-profile experiment. Opponents, meanwhile, will not want to overstate the allegedly dire consequences of a policy, lest they inadvertently lower the implicit threshold for judging the experiment's success.

With self-execution, this implicit threshold becomes explicit, an agreed-upon benchmark for assessing an experiment's success. Negotiation on such a benchmark could increase the goal transparency of experimentation, because policymakers are forced not only to identify a policy, but also to specify a basis for measurement and a threshold that would count as success. Once again, partisans would confront tradeoffs. For proponents, insistence on too low a benchmark of

[52] I do not mean to claim that policymakers will never find bases for arguing in favor of continuing experimental policies that appear to have failed. For example, in the recent Medicare experiments, after preliminary data suggested that the experiment might fail, some legislators cautioned against cancellation of the experiment on the ground that those who had benefited from the experiment might suffer if it were discontinued. *See* Abelson, *supra* note 13.

success would signal that the proponents do not believe that the policy has a good chance of achieving greater success. For a job training assistance experiment, if proponents claim that it would be a success if there is a one percent increase in employment in the treatment group relative to the control group, the opponents might respond that the job training is too expensive to justify such an increase. Promising too much, however, risks the experiment failing. Similarly, for opponents of a policy, insisting on too great a benefit might seem to concede that an experiment is expected to produce some benefit.

This analysis assumes that a self-executing experiment would be focused on just a single policy dimension (such as increased employment). It would also be possible, however, for the criterion of success to depend on some formula aggregating multiple dimensions. For example, if there is debate about how much job retraining would cost, policymakers might agree that the experiment should be viewed as a success if each increased job costs the government no more than a set amount of money. By creating an easily understood cost measure that incorporates multiple dimensions of a policy problem, policymakers could further increase goal transparency. It would also be possible to create more elaborate formulas that take into account a larger number of criteria, but past a certain point, increased complexity, however desirable from a policy standpoint, may make debate on experimentation less transparent.

However numerous the measurements of a policy's effect, a self-executing experiment need not necessarily focus on a narrow proposal. Altering only one policy variable may allow for a more scientific assessment, isolating that variable from other policy dimensions. Nonetheless, testing an entire, multi-dimensional policy may make more sense from a policy perspective. Within the policy process, it may matter not so much why a particular proposal worked (though such information would be helpful for later policies), but whether it worked according to the agreed-on criteria. A self-executing experiment that directly tests a policy that some policymakers in fact wish to implement may be most useful from the perspective of goal transparency, even when it may be difficult to understand which of many aspects of the policy ultimately were responsible for any beneficial effects. An experimental test of an entire policy makes it difficult for those disappointed by the experimental results to dismiss those results as irrelevant to actual policy or attributable to an artificial experimental environment.

Indeed, the most promising self-executing experiments might be bidirectional ones that offer two treatment groups (perhaps with a control group representing the status quo policy as well), rather than a treatment group and a single control group. Suppose, for example, that different legislators have opposite views about securities disclosure regulation, with some arguing that more stringent regulations would benefit shareholders,[53] and others insisting that excessive disclosure regulation harms shareholders.[54] A randomized experiment could randomly assign all the regulated corporations to either the "more stringent" or "more relaxed" testing group, whether or not some randomly selected corporations continue to operate under the status quo regulations. The policymakers might agree that after some evaluation period,[55] the two treatment groups will be compared, for example by assessing the average stock price increase in the two groups.[56] If one group outperformed the other by at least a set margin, then the law assigned to that group could automatically be extended to apply to all other firms.[57]

B. Possibilities for Institutional Realization

The benefits of experimental government may be more than the sum of the benefits of individual experiments. Rather, the existence of an experimental legal culture—that is, of a government in which experiments are a common feature of the policymaking process—could focus debate on the empirical effects of policies and constrain policymakers who might make unsupportable empirical claims. An experimental legal culture, however, seems likely to emerge only if legal experimentation gradually becomes more common, building on legal experiments

---

[53] *See, e.g.*, Nicholas L. Georgakopoulos, *Why Should Disclosure Rules Subsidize Informed Traders?*, 16 INT'L REV. L. & ECON. 417 (1996) (defending disclosure rules on the ground that they subsidize informed trading and therefore improve capital market liquidity).

[54] *See, e.g.*, Roberta Romano, *Empowering Investors: A Market Approach to Securities Regulation*, 107 YALE L.J. 2359, 2373-81 (1998) (providing an empirical critique of federal securities disclosure regulation).

[55] The evaluation period would need to be long enough for the difference in regulation to make a sufficient difference. Given the existing disagreement about which disclosure regime will maximize share price, there also might not be agreement about how long would be needed for stock price to be impacted. Of course, an experiment that reveals that the stock price differences between the two groups is small would itself be useful.

[56] It would also be possible to take into account other variables, such as the variance in stock price over time.

[57] A slightly different approach would be for the selected legal regime to be extended immediately only to the firms still governed by the status quo, with the firms receiving the lesser treatment required to follow that treatment for some additional years. There are at least two possible justifications for this approach. First, there might be high transition costs in switching regulatory regimes. *See generally* Michael P. Van Alstine, *The Costs of Legal Change*, 49 UCLA L. REV. 789 (2002) (arguing that transition costs must be considered in making decisions about whether to adopt new legislation). Second, the treatment will have a greater impact on stock price if it is expected to endure. With a temporary treatment, the stock price will reflect anticipation that the treatment will be removed, and thus the only stock price effects will be from changed disclosure practices during the treatment period.

credited with providing valuable information. This Part assesses the potential virtues of an experimental legal culture in particular policymaking institutions, while also considering shorter term feasibility. A legislative experimental culture might be especially useful, but development of such a culture seems more likely to emerge in administrative agencies. The view from the Progressive era that administrative agencies can and should approach policy in a scientific way retains some political attraction.[58] Courts could benefit from experimentation as well, but such experimentation seems less appropriate when initiated by individual judges than by bodies overseeing the judiciary.

1. Legislatures

The benefits of self-execution and randomization would be at their apex with legislative bodies. The past history of experimentation suggests that legislatures ordinarily pay little attention to experimental results,[59] so randomization may be useful as a way of ensuring that information is distilled in an easily digestible and relatively incontrovertible form. Self-execution, meanwhile, may be necessary as a means of ensuring that policies can be enacted into law. Because legislatures will generally reflect a broad range of ideological views, which may correlate with predictions about the empirical effects of policies, self-executing randomized experiments can be designed in a way that makes those on both sides of an issue believe that they will ultimately benefit, thus helping to facilitate passage of legislation. And finally, because legislators are electorally accountable to the public, the goal transparency of experimentation might be both particularly useful and effective.

Yet a culture of experimentation may be particularly hard to achieve in legislative bodies. Such bodies are procedurally conservative.[60] Backlash against randomization and self-execution may be particularly powerful in the legislative context. If constituents believe that legal randomization is inherently unfair,[61] legislators will likely resist it, or at least will want any

---

[58] *See generally* Marshall J. Breger & Gary J. Edles, *Established by Practice: The Theory and Operation of Independent Federal Agencies*, 52 ADMIN. L. REV. 1111, 1131-32 (2000) (providing a historical overview of Progressive era views of administrative agencies).

[59] *See supra* note 24 and accompanying text.

[60] Rules of legislative procedure do evolve, but slowly. *See, e.g.*, Saul Levmore, *Parliamentary Law, Majority Decision Making, and the Voting Paradox*, 75 VA. L. REV. 971, 976-78 (1989) (tracing the evolution of parliamentary procedure in the United States).

[61] *Cf.* Richard Arnott & Joseph E. Stiglitz, *Randomization with Asymmetric Information*, 19 RAND J. ECON. 344, 360 (1988) (noting that lotteries are generally viewed as fair in some situations, such as a military draft, and unfair in others, such as where

concrete experiments to emanate from administrative agencies rather than the legislature, to defuse responsibility. For an experimental legislative culture to develop, society generally must gain more comfort with randomization. Conceivably, the private sector's increased reliance in recent years on randomized experimentation[62] could help, though such practices could anger consumers and produce a backlash.[63] A more plausible road to public acceptance would be high-profile random experiments initiated by administrative agencies.

2. Administrative Agencies

The current Medicare experimentation for improving chronic care illustrates both the promise of administrative agency experimentation and its limits. The statute provides "for the development, testing, and evaluation of chronic care improvement programs using randomized controlled trials."[64] This is, admittedly, the most familiar form of experimental randomization, applying to consenting[65] individuals undergoing medical treatment. The test, however, is of a governmental program, rather than of a drug. The statute, moreover, provides for self-execution in the form of an independent evaluation of the randomized controlled trials,[66] specifying that if the agency finds the program to be successful according to specified criteria,[67] it shall expand the program "to additional geographic areas," possibly "includ[ing] the implementation of the program on a national basis."[68]

It should not be surprising that this experimentation is located within an administrative agency. Whether or not legislatures are well-equipped to specify technical details such as those of experimental protocols, legislatures have traditionally left such details to administrative agencies. Moreover, a defense of delegation is agency expertise, and random experimentation is a familiar tool of scientific experts. Finally, agencies routinely have rulemaking power, enabling

---

randomization produces horizontal inequity).

[62] *See generally* IAN AYRES, SUPER CRUNCHERS: WHY THINKING-BY-NUMBERS IS THE NEW WAY TO BE SMART 46-63 (2007).

[63] For example, consumers sometimes resent it when online retailers charge different consumers different prices. *See, e.g.*, Robert M. Weiss & Ajay K. Mehrotra, *Online Dynamic Pricing: Efficiency, Equity and the Future of E-Commerce*, 6 VA. J.L. & TECH. 11, 11 (2001) (discussing consumer reaction to dynamic pricing by Amazon.com).

[64] 42 U.S.C. § 1395b-8(b)(1).

[65] The statute does not specifically require that consent be obtained, but the agency administering the experiments limited the program to consenting patients. *See* MCCALL ET AL., *supra* note 21, at 22.

[66] 42 U.S.C. § 1395b-8(b)(6).

[67] *Id.* § 1395b-8(c)(2).

[68] *Id.* § 1395-b-8(c)(1).

them to adjust their programs and procedures, incorporating new information from any source. The legislative instruction that the agency consider the experimental results reinforces what the agency ought do in any event. In short, such experimentation fits within existing conceptions of the administrative function.

The Medicare experiment might itself deserve to be counted as a success even if the underlying program is labeled a failure; showing that a program fails is as important as showing that one succeeds. The experiment's benchmark for the program, however, is not an ideal model for future experiments. A program can be successful only if it "improve[s] the clinical quality of care," "improve[s] beneficiary satisfaction," and meets goals of "budget neutrality."[69] That is, a program would count as a failure if it improved care, but cost money rather than saves money. The statute does not provide for a cost-benefit analysis that would allow all improvements to care that are cost-justified. Nor does it test existing programs to determine whether any benefits that they provide are in fact cost-justified.[70] It may be easier for a legislature to require an administrative agency to implement experimentation in search of a free lunch than to use experimentation to explore program tradeoffs.

Another limitation of the Medicare experimentation is that it is limited to a specific area of concern within the Medicare program, the improvement of chronic care. More ambitious legislation would require Medicare to experiment systematically with every aspect of care, including financial details such as payments to physicians[71] and coinsurance requirements.[72] Yet another level of ambition would be to experiment with legal issues such as malpractice[73] or to lower or raise the extent of privatization in Medicare.[74] Perhaps successes could lead to broader scale use of experimentation, but these successes would need to trigger a rethinking of the function of an administrative agency, from an entity that applies existing knowledge to one that systematically generates knowledge in the face of empirical uncertainty.

---

[69] 42 U.S.C. § 1395b-8(c)(2).

[70] A concern about such a test would be that existing program participants will not likely consent to participating in an experiment in which they might receive fewer benefits. *See infra* Part IV.B.1.a (assessing whether consent ought to be required).

[71] *See, e.g.*, Thomas Bodenheimer et al., *Can Money Buy Quality? Physician Response to Pay for Performance*, 3 IND. HEALTH L. REV. 445 (2006) (assess plans for Medicare and private insurers to tie physician reimbursement to measures of performance).

[72] *See, e.g.*, William C. Hsiao & Nancy L. Kelly, *Medicare Benefits: A Reassessment*, 62 MILBANK Q. 207, 212 (1984) (assessing the case for tying Medicare coinsurance rates to beneficiary income).

[73] *See, e.g.*, William M. Sage, *The Role of Medicare in Medical Malpractice Reform*, 9 J. HEALTH CARE L. & POL'Y 217 (2006).

[74] *See, e.g.*, Gillian E. Metzger, *Privatization as Delegation*, 103 COLUM. L. REV. 1367, 1380-83 (2003) (discussing the current extent of privatization of Medicare).

Another reason that this experiment does not necessarily signal a new era of experimental government is that the legislature, rather than the agency itself, initiated the experimental process.[75] Although legislative interest in experimentation is promising, a randomized experiment initiated by an agency in the absence of an explicit legislative instruction seems more likely to move agencies' culture in an experimental direction. Organic statutes do not generally prohibit agency experimentation. Although the statutes governing the SEC, for example, do not explicitly provide for the agency to engage in experimentation, they do not prohibit experiments either.[76] Given its broad regulatory authority,[77] the SEC should be able to implement an experiment randomized at the firm level. If such an action were challenged, the courts would presumably examine it with the usual tools of administrative judicial review, ensuring for example that the action was procedurally proper,[78] was consistent with law,[79] and represented a permissible policy judgment.[80]

These hurdles should be straightforward for an agency to clear. As long as an agency goes through the ordinary notice-and-comment process,[81] providing a detailed explanation of the purpose of an experiment in the notice of proposed rulemaking,[82] as well as a "concise, general statement" of basis and purpose,[83] there should be no procedural obstacle to proceeding with an experiment that would change the law for certain entities. As long as neither the experimental

---

[75] The experimentation did stem from nonrandomized demonstration programs conducted by the Centers for Medicare & Medicaid Services. *See* Medicare Prescription Drug, Improvement, and Modernization Act of 2003 House Conference Report No. 108-391, Nov. 21, 2003, at 2083. But the fact that the agency proceeded with randomized experimentation only after legislative authorization shows that agency officials believed either that such authorization was required, or at least that it would be prudent to seek such authorization before conducting experiments.

[76] Indeed, the SEC recently engaged in a nonrandomized experiment by temporarily changing rules governing short selling. *See* Jenny Anderson, *S.E.C. Unveils Measures to Limit Short-Selling*, N.Y. TIMES, July 16, 2008, at C1. Unsurprisingly, given the absence of a randomized control group, the experiment's results are not easy to assess. *See* Floyd Norris, *Did It Help to Curb Short Sales?*, N.Y. TIMES, Aug. 13, 2008, at C1.

[77] *See, e.g.*, 15 U.S.C. § 10(b) (giving the SEC power to define "any manipulative or deceptive or contrivance").

[78] *See, e.g.*, 5 U.S.C. § 553 (2006) (setting forth requirements for notice-and-comment rulemaking).

[79] *See, e.g.*, Chevron U.S.A., Inc. v. Natural Resources Defense Council, Inc., 467 U.S. 837 (1984) (setting forth the modern standard for evaluating agency legal interpretations).

[80] *See, e.g.*, 5 U.S.C. § 706(2)(A) (2006) (requiring the reviewing court to "hold unlawful and set aside agency action . . . found to be . . . arbitrary, capricious, an abuse of discretion, or otherwise not in accordance with law").

[81] *Id.* § 553(b) (requiring publication of general notice of proposed rulemaking); *id.* § 553(c) (requiring opportunity for comment and issuance of concise general statement of basis and purpose).

[82] Agencies generally seek to meet the "general notice" requirement by publishing the actual rules that they are considering implemented, though even this is sometimes inadequate. *See, e.g.*, Portland Cement Ass'n v. Ruckelshaus, 486 F.2d 375 (D.C. Cir. 1973) (vacating a rulemaking for failure to release background documents necessary to be able to respond to notice).

[83] "Concise" and "general" are sometimes interpreted to meet "detailed" and "specific." *See* Automotive Parts & Accessories Ass'n v. Boyd, 407 F.2d 330, 338 (D.C. Cir. 1968) (warning against interpreting these words literally).

nor the control legal regimes is inconsistent with the SEC's governing statute,[84] a decision to launch an experiment should present no problem for *Chevron* review.[85] Perhaps the most significant obstacle would be hard look review,[86] in which a court would examine the agency's justification for creating the experiment. But hard look review is supposed to be deferential,[87] and an agency should be able to justify employing a randomized experiment on the ground that this approach could provide information relevant to the administrative process.

Indeed, an administrative agency should perhaps receive broader latitude to create an experiment than to create a new administrative regime without an experiment. Procedurally, an agency might argue that it should not have to go through notice-and-comment to establish an experiment,[88] because the experiment is merely designed to produce data from which to make a subsequent policy decision. Courts have been hesitant to allow agencies to avoid the notice-and-comment process for temporary rules,[89] perhaps in part because this would allow an administrative agency to renew a program indefinitely.[90] An agency should, however, at least be allowed to focus solely on the reason for conducting an experiment, rather than responding to comments on the merits of the underlying policy issue. Because an experiment produces data on

---

[84] Given the vagueness of the statutes governing the SEC, this seems unlikely. *See, e.g.*, 15 U.S.C. § 78j (2006) (making it unlawful for any person "[t]o use or employ . . . any manipulative or deceptive device or contrivance in contravention of such rules and regulations as the Commission may prescribe as necessary or appropriate in the public interest or for the protection of investors"); *id.* § 78*l*(giving the Commission authority to require disclosure of information "as necessary or appropriate in the public interest or for the protection of investors").

[85] Chevron U.S.A., Inc. v. Natural Resources Defense Council, Inc., 467 U.S. 837 (1984).

[86] *See, e.g.*, Greater Boston Television Corp. v. FCC, 444 F.2d 841, 851 (D.C. Cir. 1970) ("Its supervisory function calls on the court to intervene . . . if the court becomes aware . . . that the agency has not really taken a 'hard look' at the salient problems, and has not genuinely engaged in reasoned decisionmaking."); *infra* notes 95-99 (discussing the seminal Supreme Court hard look case).

[87] *See, e.g.*, Motor Vehicle Manufacturers Ass'n v. State Farm Mutual Automobile Ins. Co., 463 U.S. 29, 43 (1983) ("The scope of review under the 'arbitrary and capricious' standard is narrow and a court is not to substitute its judgment for that of the agence.").

[88] This might be so when Congress has explicitly instructed an agency to conduct an experiment. Even here, however, the agency is generally like to notify the public of its intent to run an experiment, for example by soliciting third parties to perform the experiments. *See, e.g.*, *Medicare Program; Solicitation for Proposals for the Demonstration Project for Disease Management for Severely Chronically Ill Medicare Beneficiaries with Congestive Heart Failure, Diabetes, and Coronary Heart Disease*, 67 FED. REG. 8267-001, 2002 WL 245888.

[89] The Administrative Procedure Act continues a general exemption for notice-and-comment "when the agency for good cause finds . . . that notice and public procedure thereon are impracticable, unnecessary, or contrary to the public interest"). 5 U.S.C. § 553(b)(B). But the courts have found that the temporary nature of a rule is not enough to escape notice-and-comment. *See* Tenn. Gas Pipeline v. FERC, 969 F.2d 1141 (D.C. Cir. 1992).

[90] *But see* Juan J. Lavilla, *The Good Cause Exemption to Notice and Comment Rulemaking Requirements Under the Administrative Procedure Act*, 3 ADMIN. L.J. 317, 378 (1989) (suggesting that courts enforce temporary rules only until the agency has had enough time to develop a permanent rule, whether or not it has done so).

a policy issue, courts should not require an agency to show that existing data already justifies the policy that the experiment is designed to test.

To enact an experimental policy permanently, an agency presumably would face hard look review, but here too the courts should perhaps be more deferential than usual. Critics of notice-and-comment have complained that it is "ossified,"[91] making it too cumbersome to effect change. A response to this objection is that demanding review by the courts ensures that an agency does not pursue an idiosyncratic, ideological agenda.[92] An agency conducting an experiment is less likely to be doing so than an agency drawing inferences based on existing data that plausibly might support different conclusions. Moreover, courts conducting judicial review should recognize the unique value of evidence from randomized experimentation.[93] There remains a danger that an agency might make invalid inferences on the basis of an experiment. At least where experiments provide the best available evidence on a policy issue, however, courts should allow an agency to reply that it placed more weight on the experimental evidence, without chronicling on a case-by-case basis all of the problems of nonexperimental evidence.

If courts did accord data from random experiments more weight than policy speculation or other data, agencies might engage in experimentation to constrain judges who will review the new policy. Agencies, after all, are not the only possible source of ideological influence in administrative law; studies repeatedly show that political variables help explain judges' decisions in performing judicial review of administrative action.[94] If a random experiment provides evidence that a policy advances some important objective, then it will be more difficult for judges to vacate a rulemaking implementing the new policy on judicial review. Judges do not generally second-guess an agency's priorities, but do scrutinize empirical evidence.

Consider, for example, the Supreme Court's seminal decision in *Motor Vehicle Manufacturers Ass'n v. State Farm Mutual Automobile Insurance Co.*,[95] in which the Court

---

[91] *See, e.g.*, Thomas O. McGarity, *Some Thoughts on "Deossifying" the Rulemaking Process*, 41 DUKE L.J. 1385 (1992); Richard J. Pierce, Jr., *Seven Ways to Deossify Agency Rulemaking*, 47 ADMIN. L. REV. 59 (1995).

[92] An ideological agency, facing an ideologically hostile court, might respond either by investing more in meeting the requirements of the hard look doctrine or "allocate their resources to other projects." Richard L. Revesz, *Environmental Regulation, Ideology, and the D.C. Circuit*, 83 VA. L. REV. 1717, 1770 (1997).

[93] *Cf.* Dorf & Sabel, *supra* note 1, at 397 (arguing that hard look review should reward experimental agency approaches, though not focusing specifically on random experimentation).

[94] *See, e.g.*, Thomas J. Miles & Cass R. Sunstein, *Do Judges Make Regulatory Policy? An Empirical Analysis of* Chevron, 73 U. CHI. L. REV. 823 (2006) (offering a comprehensive analysis of Supreme Court and circuit court review of agency legal conclusions).

[95] 463 U.S. 29 (1983).

struck down the National Highway Traffic Safety Administration's cancellation of a standard requiring installation of either air bags or passive seat belts in cars.[96] Part of the Court's concern was that the agency could not substantiate its claim that the requirement had led to at least a five percent increase in seat belt use.[97] Suppose, however, that the agency had run a randomized experiment assessing the effectiveness of such a system.[98] Then the agency would have had relatively incontrovertible data about the increase in seat belt use. To overturn the agency's decision on this ground,[99] the Court would have had to question the agency's normative judgment about the value of whatever level of increased seat belt use was demonstrated. But, under current understandings of the hard look doctrine stemming from the *Motor Vehicle Manufacturers* decision, courts applying the hard look doctrine do not substitute their value choices for an agency's.

When an agency adopts a self-executing experiment, a court reviewing the policy should assess it from an ex ante rather than an ex post perspective.[100] It may have been sensible for an agency to specify that policy will depend on a simple numerical comparison of the results of an experiment even if statistical arguments based on the experimental data could support both positions. This presents a tradeoff between information and objectivity. Even if hypothetical neutral observers could come closer to making sound empirical judgments by second-guessing experimental results, allowing judges the power to do so entails the risk that the judges may strike down agency policies based on their own normative preferences. Where an agency reasonably commits itself to a self-executing experiment, especially if opponents of a policy do not generally complain at the time an experiment is initially designed, a court should assess that

---

[96] *Id.* at 34.

[97] *Id.* at 53-54.

[98] For example, the agency might have required automobile manufacturers to install passive seat belts in a randomly selected proportion of cars ordered by consumers direct from the factory, and then tracked accident data for these cars. This research design is not without its problems; some consumers might dislike the passive belts and return the cars for alternatives. But part of the agency's concern was that consumers might disable the passive seatbelt function. *Id.* at 54.. Thus, an "intent to treat" estimate, *see infra* note 221 and accompanying text, should at least approximately capture the effect of offering the feature by default at no extra charge.

[99] The Court had an alternative basis for overturning the agency's decision, that the agency did not consider requiring air bags. *Id.* at 48-50. Of course, the agency also might have run a randomized study of the effectiveness of air bags.

[100] Admittedly, it may not be easy for judges to overcome hindsight bias. Jeffrey J. Rachlinski, *A Positive Psychological Theory of Judging in Hindsight*, 65 U. CHI. L. REV. 571, 588-94 (1998). Courts might, however, consider challenges to a self-executing experiment before the experiment is complete.

decision on the basis of the information then available,[101] not on the basis of information that becomes available after the decision is made.

Administrative law doctrine, in short, can support an experimental culture by placing great weight on the results of randomized experiments in performing judicial review, and by supporting efforts of agencies to generate support among opponents for experiments that would resolve their differences. A more aggressive use of administrative law in favor of experimentation would be for courts to strike down policies on the basis that an administrative agency failed to conduct a randomized experiment, when such an experiment would have resolved key empirical questions. This would require no great doctrinal innovation, as courts already sometimes fault agencies for failing to gather relevant data.[102] It seems unlikely that courts will insist on experimentation until experimentation becomes commonplace, but ultimately such insistence can be justified on the ground that agencies ought to use the policymaking tools that best resolve the policy issues facing them.

3. Courts

The third branch of government is generally least suited to initiate randomized experiments. Judges decide individual cases, at least theoretically resolving broader legal issues only to the extent necessary to resolve the cases before them. Perhaps the traditional view that "flipping a coin" is the antithesis of judging could be relaxed when randomization is for the purpose of conducting an experiment.[103] But an individual judge cannot typically initiate a randomized experiment anyway, because a judge will not preside over a sufficiently large number of cases to make any randomization meaningful. Nonetheless, with legislative authorization, or with delegation to a quasi-judicial body,[104] experiments could randomize different cases to different treatments.[105] For example, the literature on fee shifting focuses on

---

[101] Courts conversely may uphold an agency's decision only on the basis of grounds that the agency in fact relied upon. *See* SEC v. Chenery, 318 U.S. 80, 87 (1943) ("The grounds upon which an administrative order must be judged are those upon which the record discloses that its action was based.").

[102] *See, e.g.*, Portland Cement Ass'n v. Ruckelshaus, 486 F.2d 375 (D.C. Cir. 1973).

[103] *See, e.g.*, Richard K. Greenstein, *Why the Rule of Law?*, 66 LA. L. REV. 63, 89-90 (2005) (discussing the problems of making decisions by flipping coins).

[104] *See* Thomas E. Willging, *Past and Potential Uses of Empirical Research in Civil Rulemaking*, 77 NOTRE DAME L. REV. 1121 , 1204 (2002) (suggesting that the Standing Committee on Rules of Practice and Procedure of the Judicial Conference of the United States conduct experiments).

[105] Laurens Walker has argued for the use of field experiments to test and improve the *Federal Rules of Civil Procedure*. *See* Walker, *supra* note 5.

laboratory experiments[106] and nonexperimental statistical analyses.[107] A randomized experiment could potentially resolve the theoretical controversy over whether fee shifting would increase or reduce litigation costs,[108] though it could be harder to measure effects of fee shifting on variables such as adjudication accuracy.[109] Individual jurisdictions sometimes innovate in case management and processing,[110] and randomized experiments could help assess these innovations.[111]

C. Alternative and Related Approaches

Experimentation is not the exclusive means by which legislatures can improve information, make enactments of reform more likely, and enhance transparency. This section considers a variety of other available techniques—sunset clauses, reporting requirements, delegation, and privatization—and assesses their advantages and disadvantages relative to experimentation, as well as how these techniques can be used in conjunction with randomized self-executing experiments.

1. Sunset Clauses

Sunset clauses are legislative provisions that cause legislative programs to expire, unless the legislature acts to renew those programs.[112] Sunset clauses can provide some of the same

---

[106] *See* Laura Inglis et al., *Experiments on the Effects of Cost-Shifting, Court Costs, and Discovery on the Efficient Settlement of Tort Claims*, 33 FLA. ST. U. L. REV. 90 (2005).

[107] *See, e.g.*, Jackson Williams, *Effects of Attorney Fee Shifting Laws on Claiming Behavior*, 34 POL'Y SCIENCES 347, 354 (2001) (finding that fee shifting rules increase the incidence of litigation, but noting that endogeneity could drive the result).

[108] *See generally* Steven Shavell, *Suit, Settlement, and Trial: A Theoretical Analysis under Alternative Methods for Allocation of Legal* Costs, 11 J. LEGAL STUD. 55, 69 (1982) (showing that such effects depend in part on empirical factors such as whether plaintiffs' chances of prevailing are generally low or high).

[109] *See* Keith Hylton, *Fee Shifting and Incentives to Comply with the Law*, 46 VAND. L. REV. 1069, 1071 ("Not a shred of empirical evidence on the compliance effects of alternative fee shifting rules exists, however, and it is unlikely that it ever will, given the cost of the required experiments.").

[110] *See, e.g.*, Harry N. Scheiber, *Innovation, Resistance, and Change: A History of Judicial Reform and the California Courts, 1960-1990*, 66 S. CAL. L. REV. 2049, 2065 (1993) (discussing procedural innovation by state and federal district courts in California).

[111] An example of nonrandomized experiments that are difficult to analyze is the advent of "drug courts." *See* Morris B. Hoffman, *The Drug Court Scandal*, 78 N.C. L. REV. 1438, 1479-80 (2000) (noting the difficulties of interpreting nonrandomized experiments on drug courts); *cf.* Denise C. Gottfredson & M. Lyn Exum, *The Baltimore City Drug Treatment Court: One-Year Results From a Randomized Study*, 39 J. RES. CRIME & DELINQ. 337, 341 (2002) (providing a randomized study of a drug court and noting that the many previous studies were nonrandomized).

[112] Jacob Gersen defines sunset clauses as a subset of a broader class of "temporary legislation," with sunset clauses specifically focused on programs run by administrative agencies. *See* Jacob E. Gersen, *Temporary Legislation*, 74 U. CHI. L. REV. 247, 260 (2007) (noting that sunset legislation has "sought to increase legislative oversight, bureaucratic responsiveness, and regulatory efficiency"). Here, I use the "sunset clauses" phrase synonymously with "temporary legislation."

benefits as random experimentation. Sunset clauses test particular legal reforms, providing the legislature additional information when it considers renewing the program.[113] Legislation can always be changed by legislatures, but sunset clauses increase the probability of later legislative scrutiny, making it more likely that the legislature will consider information produced by the experiment. Sunset clauses also can be easier to enact, because marginal legislators may be reassured that they will have an opportunity to revisit the issues later with better information.[114] Finally, by forcing legislators periodically to reexamine issues, they make it more difficult for legislators to escape accountability by inaction, thus increasing goal transparency.

There are, however, significant differences between the approaches. Some of these considerations favor self-executed randomized experiments. Where feasible, randomized experiments may produce more reliable information.[115] Self-execution ensures that the legislature will take the results of the experiment into account. In addition, self-execution requires legislators to set goals explicitly, while agreement to a sunset clause represents a loose commitment to revisiting an issue. Meanwhile, self-execution saves legislators the time that a sunset clause would require to re-enact a statute.[116] Other considerations, however, may favor sunset clauses. If empirical uncertainty exists about whether a program will be politically acceptable, a sunset clause may be effective.[117] It might be difficult to conceive of an experiment that could compare public reaction to the treatment and control regimes. Meanwhile, if legislators believe that it will be particularly important for public attention to be drawn again to an issue, the sunsetting of the law might provoke more public attention than self-execution of an experiment.

Randomization can easily be used in conjunction with a sunset clause. Consider, for example, a current proposal that would sometimes channel patent cases to judges with more

---

[113] *See id.* at 274 ("Experimental temporary legislation tends to implement policy on a short-term basis as a means of generating information that can be subsequently incorporated into the policymaking process.").

[114] *Id.* at 264 ("[T]emporary legislation's initial enactment costs are almost certainly less than permanent legislation's.").

[115] *See infra* Part III.A.

[116] Gersen notes that whether sunset clauses save legislative time overall is ambiguous. *See* Gersen, *supra* note 112, at 263-66. He notes, though, that especially for legislation of relatively short duration (such as three years), sunset clauses are likely to be more demanding on legislative time. *Id.* at 264. Conceivably, an agreement to a self-executing experiment might increase or decrease legislative time relative to permanent legislation. On one hand, legislators will need to define (or delegate the task of deciding) how the experiment will operate, but on the other, a self-executing experiment may allow legislators an escape from the necessity of resolving their disagreements through negotiation.

[117] Sometimes, however, even politically controversial provisions are extended. *See, e.g.*, Heather Hillary & Nancy Kubasek, *The Remaining Perils of the Patriot Act: A Primer*, 8 J.L. SOC'Y 1, 26 (2007) (noting the extension of controversial sunsetting provisions of the USA Patriot Act).

patent experience.[118] Congress could provide for a randomized experiment of this proposal over a limited time,[119] and could specify a simple metric for assessing the success of the experiment.[120] Merely by virtue of advance specification, this metric would presumably receive attention at the time reauthorization is considered. A self-executing experiment that provides for the legal regime to be continued and extended only given some metric of success in effect sunsets the legislation in the absence of success.

2. Reporting Requirements

Another tool that increases the chance of reconsideration is a requirement that a report evaluating a program be prepared, either by the implementing agency or an independent agency, such as the Government Accounting Office.[121] The report might set forth specific recommendations about how the law should be changed. Such a reporting requirement can increase information available to legislators, in much the same way as an experiment, and the expectation of such new information might mollify wavering legislators. In addition, such a requirement can promote transparency; legislators may be hesitant to approve legislation that a later report is likely to pillory.

These effects seem likely to be considerably weaker than self-executing randomized experiments'. Many governmental reports receive little public notice,[122] and like experiments that are not self-executing, may not influence legislators at all. A report may have some advantages over a self-executing experiment, however. It can consider all aspects of experience with a new legal program, including aspects that either are not susceptible to experimentation, or that designers of a self-executing experiment simply might not have anticipated. Once again, a hypothetical neutral observer might sometimes prefer a report to the results of a simplified

---

[118] *See* H.R. 34 (110th Cong. 2007). For an empirical analysis, see David L. Schwartz, *Practice Makes Perfect? An Empirical Study of Claim Construction Reversal Rates in Patent Cases*, 107 MICH. L. REV. (forthcoming 2008).

[119] Under the proposal, judges who do not opt in to being specialized patent judges are permitted to refuse patent cases, in which case they are reassigned to judges who have opted in. H.R. 34 § 1(a)(1)(C)-(D). A randomization could provide that a decision to opt out would only have a 50% chance of being granted.

[120] For example, Congress could compare reversal rates in cases in which the proposal applied to reversal rates in cases in which the proposal did not apply.

[121] *See, e.g.*, Pub. L. No. 106-65, § 804(d), 113 Stat. 512, 704 (1999) (requiring the GAO to issue a report one year following the enactment of certain regulations).

[122] *See, e.g.*, Nicholas A. Robinson, *Legal Systems, Decisionmaking, and the Science of Earth's Systems: Procedural Missing Links*, 27 ECOLOGY L.Q. 1077, 1105 (2005) (noting that even national environmental reports required to be filed with the United Nations are largely ignored).

randomized experiment, but the simplification inherent in an experiment and in self-execution may increase the chance that information has an effect on the policy process.

A special case of a reporting requirement is the creation of a "blue-ribbon" panel, typically consisting of an illustrious bipartisan group. In recent years, such commissions have addressed issues such as preventing further terrorist attacks,[123] the failure of intelligence on whether Iraq possessed weapons of mass destruction,[124] the further conduct of the war in Iraq,[125] voter identification requirements,[126] and income tax reform.[127] To the extent that such panels can reach consensus, the public receives assurance that its recommendations are apolitical.[128] Such reports may produce information, increase the chance that reform is enacted,[129] and expose legislative positions in tension with available evidence. Both experiments and blue-ribbon panels are means by which a legislature, unable to agree itself on issues, can provide a relatively objective type of test of a policy issue. Which is preferable may depend on the amenability of the issue to experimentation and on the chance that the panel will successfully influence the policy process.

3. Delegation to Independent Agencies

A blue ribbon panel can be conceived as roughly equivalent to an independent agency, but with power only to make recommendations.[130] Like a self-executing experiment, an independent agency is a mechanism for gathering information and changing the law without

---

[123] *See* NAT'L COMM'N ON TERRORIST ATTACKS UPON THE U.S., THE 9/11 COMMISSION REPORT (2004), http://www.gpoaccess.gov/911/pdf/fullreport.pdf.

[124] *See* COMM'N ON THE INTELLIGENCE CAPABILITIES OF THE U.S. REGARDING WEAPONS OF MASS DESTRUCTION, REPORT TO THE PRESIDENT OF THE UNITED STATES (2005), http://www.wmd.gov/report/wmd_report.pdf.

[125] *See* THE IRAQ STUDY GROUP, THE IRAQ STUDY GROUP REPORT: THE WAY FORWARD – A NEW APPROACH (2007), *available at* http://www.usip.org/isg/iraq_study_group_report/report/1206/index.html.

[126] *See* COMM'N ON FEDERAL ELECTION REFORM, BUILDING CONFIDENCE IN U.S. ELECTIONS (2005), *available at* http://www.american.edu/ia/cfer/report/full_report.pdf.

[127] *See, e.g.*, PRESIDENT'S ADVISORY PANEL ON FEDERAL TAX REFORM, FINAL REPORT (2005), *available at* http://www.taxreformpanel.gov/final-report/.

[128] That does not mean that recommendations are always unanimous and uncontroversial. *See, e.g.*, Spencer Overton, *Carter Baker Dissent*, http://www.carterbakerdissent.com/ (offering a dissenting statement by a member of the Carter-Baker Commission, which the Commission as a whole did not allow to be included in full in the report).

[129] This will be so, however, only if it is easier for commission members to bargain than for members of the legislature to do so. One reason that such bargaining might be more feasible is that commission members may be relatively free of the need to appease special interests, possibly reducing disagreement on some issues.

[130] Critics of delegation sometimes point out that in the absence of delegated power, independent agencies would still be able to use their expertise to develop proposals that they would then forward to the legislative branch. *See, e.g.*, Ronald J. Krotoszynski, Jr., *Reconsidering the Nondelegation Doctrine: Universal Service, the Power to Tax, and the Ratification Doctrine*, 80 IND. L.J. 239, 308-11 (2005) (arguing for a system in which Congress would be required to ratify certain agency decisions).

legislative intervention. Sometimes, delegation facilitates reform by allowing enactment of laws even where lawmakers cannot agree about some details. Common critiques of delegation, however, show advantages of self-executing random experimentation. First, delegation may be a vehicle by which legislators avoid accountability, reducing goal transparency by allowing legislators to make vague normative commitments.[131] With self-executing randomized experiments, legislators must specify the criteria that reflect whether new legislation is enacted. Second, even nominally independent agencies may reflect the political preferences and biases of the administrations that appoint their officials. By creating a randomized experiment and criteria for self-execution, a legislature, even one that relies on an agency to implement and measure the experiment, limits the degree to which the normative preferences of the executive branch will affect the law.

4. Privatization

Privatization is a form of delegation, and it is thus subject to some of the same critiques as delegation,[132] particularly that it may reduce legislative accountability.[133] Yet one approach to privatization, in which private actors are given high-powered financial incentives,[134] can enhance accountability. Suppose, for example, that the government privatizes a prison by defining a formula that promises to pay an operator based on some formula involving observable criteria, such as recidivism, escapes, inmate health, and so on, where the coefficients attached to each variable represent the approximate social benefits of improvements with regard to that consideration.[135] The government might then auction the right to operate the prison to the private operator willing to pay the most to receive the rewards of the formula.

---

[131] *See generally* George I. Lovell, *That Sick Chicken Won't Hunt: The Limits of a Judicially Enforced Non-Delegation Doctrine*, 17 CONST. COMMENTARY 79, 80-81 (2000) (assessing whether a strong nondelegation doctrine would promote accountability).

[132] *See* Gillian E. Metzger, *Privatization as Delegation*, 103 COLUM. L. REV. 1367, 1394-1400 (2003) (exploring parallels between delegation and privatization). Privatization may have some advantages relative to delegation, such as that private actors may be more efficient than public actors, as well as some costs, such as that private actors may be more prone to act in their own interest rather than in the public interest.

[133] *Id.* at 1470-80.

[134] High-powered incentives tend to work best when it is feasible to monitor performance. *See* Howard Frant, *High-Powered and Low-Powered Incentives in the Public Sector*, 6 J. PUB. ADM. RESEARCH & THEORY 365, 366-72 (1996) (applying this insight in private- and public-sector contexts).

[135] *See, e.g.*, Alexander Volokh, Privatization and the Effectiveness of Monitoring Agencies (Georgetown Law and Economics Research Paper No. 982146, April 2007), *available at* http://works.bepress.com/alexander_volokh/5 (discussing monitoring in the prison privatization context).

There may be grounds for opposing such an approach, particularly if what is delegated, as some commentators have suggested,[136] is the capacity to engage in lawmaking. Nonetheless, it could help improve goal transparency in much the same way as random experimentation, by requiring explicit specification of social welfare criteria. Meanwhile, the private organization itself would have robust incentives to develop new information to improve its processes. The difference between the two approaches is that with government-initiated random experimentation, policymakers specify criteria and the legal consequences of success, while with privatization, policymakers specify criteria but allow the private entity to determine how, within any applicable legal constraints, to optimize given those criteria. Privatization may thus increase the chance that legal structures will be adjusted to incorporate information from experimentation and elsewhere, but at the hazard of less public control.

III. ADVANTAGES AND TECHNICAL LIMITS OF RANDOMIZATION

This Part explores the advantages and limits of randomized studies. This analysis responds to two arguments. The first is that randomized studies are unnecessary, because statistical and econometric techniques can be used to estimate policy effects reliably. Part III.A argues that even when the most advanced techniques are employed, nonrandom analyses will generally leave more uncertainty than random analyses. Regression analyses will often leave sufficient ambiguity that both proponents and opponents of a particular policy can credibly disagree about whether a particular study should influence public debate. Not only are randomized studies often better than corresponding nonrandomized studies, but more importantly, the policy context magnifies the difference.

The second argument is that even if randomized studies are better than nonrandomized ones, policymakers can disagree about their relevance to policy questions. Part III.B agrees but argues that this problem strengthens the case for self-executing randomized studies. A nonrandomized study can be self-executing, with policymakers agreeing in advance on regression designs and a formula aggregating regression coefficients or other results. But it is difficult to conceive in advance of all the regression tests that would be necessary to verify robustness. Any formula is likely to be somewhat arbitrary and difficult to understand.

[136] *See* Mark A. Cohen & Paul H. Rubin, *Private Enforcement of Public Policy*, 3 YALE L.J. 167 (1985) (describing how enforcement of public policy can be privatized).

Nonrandomized statistical analyses are sufficiently unreliable that they are unlikely to influence the policy process ex post, and sufficiently opaque that they cannot serve as the basis for self-execution. Randomized statistical analyses are still somewhat unreliable as policy bases, but sufficiently good that relying on them would tend to improve policy, and transparent enough to serve as a basis for self-execution. Even if social scientists might feel more comfortable scrutinizing many nonrandom experiments than blindly following an ex ante specification of a measurement to be taken from a random experiment, randomized self-executing experiments may be the better political tool.

A. Problems with Nonrandom Evaluation

Any statutory change is experimental in that it creates a new legal regime, allowing comparison to the world in the prior regime. Indeed, it is common for proponents and neutral commentators to describe such a change as "an experiment."[137] Effects, however, can be difficult to assess, because there may be alternative explanations for any observed changes. Some legal changes are sufficiently drastic, and the responses to them sufficiently immediate and profound, that some changes may be attributed to them. But reasonable observers often disagree about causality. And even if reasonable sophisticated parties would agree, partisans may offer misleading interpretations of the data. The media may then summarize the debate by simply noting that experts disagree.[138] Those who do not have the time, inclination, or ability to probe the evidence cannot then easily discern the truth.[139]

As the number of jurisdictions trying an experiment rises, the data may become clearer. But even then, the challenges of statistical analysis may make it difficult to reach confident conclusions. Statistical associations between jurisdictions adopting policies and other variables need not imply causation. It will thus almost always be relatively easy for partisans to find some basis on which either to develop misleading results or to offer critiques of results that in fact are relatively robust. Part III.A.1 explains why even with numerous jurisdictions, conventional multiple regression analysis in which the policy of interest forms an independent variable may

---

[137] *See supra* note 1.

[138] *See* Bryan Keefer, *Tsunami*, COLUM. JOURNALISM REV., July 1, 2004, at 18 (discussing reporters' reluctance to take sides on issues of public controversy).

[139] A similar problem exists when jurors try to assess evidence beyond their competence. *See* Scott Brewer, *Scientific Expert Testimony and Intellectual Due Process*, 107 YALE L.J. 1535 (1998).

produce inaccurate results, while Part III.A.2 argues that pseudo-random experimentation techniques do not resolve the problem. These sections, of course, are not intended to provide comprehensive overviews of the uses and limits of statistical analysis.[140] Part III.A.3 comments on the difficulties of improving the law by using the states as policy laboratories.

1. Conventional Regression Analysis

a. Omitted variable bias

Correlation, introductory statistics students are told, does not imply causation. The simplest example of this is the possibility of reverse causation. For example, suppose that students who receive sex education have sex at an earlier age.[141] This could mean that sex education encourages students to have more sex, but it also could reflect that school districts with high rates of student sexual activity respond to these rates by offering sex education. A standard statistical approach to overcoming this problem is to add control variables for the characteristics of the students, such as family income, parents' education, and religion, as well as of the community, such as whether it is rural and in which region of the country it is located.[142] If those variables exhaust all nonrandom factors contributing to community and family decisions about sex education, then this technique will be successful, because the coefficient on the sex education variable then represents the effect of random variation in whether students are exposed to sex education. But if there is an omitted variable, correlated with both the community decision to offer sex education and the individual decision to have sex, the coefficient will be biased.

This problem cannot easily be avoided even by careful researchers (and can be exploited by researchers who hope to establish a particular result). There are at least two reasons for this. First, the available data may be incomplete. Even if there are strong theoretical reasons to believe that parental education is an important variable, it may be impossible to develop a measure that fully accounts for the parent's educational level.[143] For example, a measure indicating whether someone's mother graduated from high school would seem to imply that all high school dropouts

---

[140] A useful overview of regression analysis is WILLIAM MENDENHALL & TERRY L. SINCICH, A SECOND COURSE IN STATISTICS: REGRESSION ANALYSIS (6th ed. 2003). For a critical analysis of the use of empirical evidence in legal scholarship, see Lee Epstein, *The Rules of Inference*, 69 U. CHI. L. REV. 1 (2002).

[141] *See, e.g.*, Deborah Anne Dawson, *The Effects of Sex Education on Adolescent Behavior*, 18 FAMILY PLANNING PERSP. 162 (July/August 1986)

[142] *See, e.g.*, *id.* at 170 tbl. 9 (listing control variables).

[143] Dawson's study used a binary variable indicating whether the mother had at least twelve years of education. *See id.* at 166.

are alike and all high school graduates are alike, but within each group, there may be considerable educational heterogeneity. Even more precise data—including information like parental GPAs—will be at best only crude proxies. Second, the researchers' theoretical accounts of what variables may correlate with the dependent and independent variables are likely to be incomplete.

The omitted variable bias may be particularly problematic when regressions are used to analyze the behavior of individuals who have self-selected into particular governmental programs. For example, Julie Cullen et al. analyzed the effect of school choice lotteries, whose winners would be allowed to attend particular schools.[144] Students who won the school choice lotteries tended to do better than students who did not enter the lotteries. Competing explanations include that lottery winners were allowed to attend better schools and that more motivated students are likely to self-select into the lottery. In the absence of variables fully capturing student motivation, a regression would tend to inflate the apparent effects of the schools on performance. Indeed, Cullen et al. showed that students who won the school choice lotteries performed no better than students who entered but lost the same lotteries. Though not created for the purpose of facilitating data analysis, the lottery produced random assignments that allowed the researchers to avoid omitted variable bias.

Even studies that attempt to control for all available information and seek to minimize the danger of omitted variable bias may nonetheless omit important variables. This can be shown by comparing the results of randomized experiments from the results of nonrandomized statistical analysis. Paul Glewwe et al. conducted separate prospective randomized and retrospective nonrandomized studies of whether making "flip charts" available to students in Kenya increased test scores. The retrospective studies showed that flip charts increased test scores, while the randomized studies revealed no effect.[145] Even a difference-in-difference analysis gave misleading results, showing that students in schools adopting flip charts performed especially well in flip chart subjects relative to other subjects. The forces that lead jurisdictions or

[144] *See, e.g.*, Julie Berry Cullen et al., *The Effect of School Choice on Student Outcomes: Evidence from Randomized Lotteries* (NBER Working Paper No. 10113, 2003).

[145] Paul Glewwe et al., *Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya* (NBER Working Paper No. 8018, 2000).

institutions to adopt policy changes such as flip charts may be so complex that omitted variables matter even when it is not obvious that any important variables are omitted.

Other studies have also shown different results with randomized and nonrandomized studies.[146] This does not mean that retrospective statistical analysis is worthless. A rational Bayesian policymaker would update priors about likely effects of policy choices based on the regression results.[147] The degree to which priors change should depend, among other factors, on the magnitude of the effects demonstrated and on the relevance and possible implications of variables omitted from study. With large numbers of studies over time pointing largely in one direction, it may sometimes be possible for the policy community to develop a consensus on what were once controversial empirical questions.[148] But there is no single objective metric for assessing the quality of a particular study, including considerations such as how important omitted variables are or in what direction they would likely bias coefficients.

b. Publication bias and misspecification

Statistically significant results can also be spurious as a result of publication bias. A finding of a statistically significant outcome, at the generally accepted 0.05 level, means that there is a five percent chance that an outcome at least as extreme would have occurred by pure

---

[146] For example, a study by Steven Glazerman et al. evaluated the impact of job training or employment services program on future income, and compared their results with earlier, nonrandomized studies. *See* Steven Glazerman et al., *Nonexperimental Versus Experimental Estimates of Earnings Impacts*, 589 ANNALS AM. ACAD. POL. & SOC. SCI. 63 (2003). They were able to identify a variable (prior earnings) that some earlier studies had omitted, and these studies had biased estimates of about $1,600 relative to their randomized results. *See id.* at 79. Even studies that included this variable, however, had a bias of about $1,000, *see id.*, again apparently demonstrating that even unknown omitted variables may be quite important.

Another study in which randomized results seemed to contradict earlier nonrandomized analyses focused on agricultural technology adoption. *See* Esther Duflo et al., *Understanding Technology Adoption: Fertilizer in Western Kenya: Preliminary Results from Field Experiments* (2006), http://www.econ.berkeley.edu/users/webfac/saez/e231_s06/esther.pdf. Earlier studies showed that when farmers' contacts used fertilizer, the farmers were more likely to use it themselves. This supported the view that technological innovations could spread through social networks, thus making more plausible the case for government to seek to introduce agricultural technology in the hope that farmers will spread it among themselves. In the randomized study, a group of farmers were asked to list their contacts, and some of the farmers were randomly selected to receive fertilizer. These farmers' contacts were no more likely to use fertilizer than the contacts of those who did not receive it. Thus, in the nonrandomized study, there was presumably some unidentified variable that affected both whether farmers and their contacts used fertilizer, leading to the spurious conclusion that the contacts led to technology diffusion. *See also* Charles Manski, *Identification of Exogenous Social Effects: The Reflection Problem*, 60 REV. ECON. STUDS. 531 (1993) (labeling this the "reflection problem" in studies seeking to identify sociological effects among neighbors).

[147] *See supra* note 38 and accompanying text.

[148] That does not necessarily mean that meta-analyses of numerous studies can be trusted to produce reliable results, because many studies on the same issue can suffer from similar biases. Mark Lipsey and David Wilson found that substantial bias relative to the results of nonrandomized studies may still exist in individual meta-analyses of particular experimental interventions. *See* Mark W. Lipsey & David B. Wilson, *The Efficacy of Psychological, Educational, and Behavioral Treatment: Confirmation from Meta-Analysis*, 48 AM. PSYCHOL. 1181, 1193 (1993) ("In some treatment areas, therefore, nonrandom designs (relative to random) tend to strongly underestimate effects, and in others, they tend to strongly overestimate effects.").

chance if the null hypothesis were true.[149] If, for example, researchers test 100 propositions that in fact are all false and would be counterintuitive if true, about five of these tests may produce statistically significant results, and these mistaken results will be the most publishable.[150] Meanwhile, insignificant findings provide little support for the truth of the corresponding null hypotheses. Such findings also may be most publishable when they are counterintuitive, but a counterintuitive failure to reject a null hypothesis may also be the result of chance.[151]

Publication bias applies not only across studies, but also within studies. Researchers face many choices about how to specify regression equations: what functional form to use,[152] which variables to include, what transformations to apply to the variables,[153] and which observations to include.[154] Especially within social science, researchers do not necessarily settle on regression specifications in advance, but instead "pretest" data to determine which results to report.[155] Considering a large number of regression specifications may help researchers develop nuanced accounts, but researchers will generally be more likely to report results producing statistical significance.[156] Laboratory experiments are also subject to publication bias, but other researchers can attempt replication. Social science researchers cannot rerun history.[157]

Social scientists can, however, seek to assess the robustness of published results. A recent example was John Donohue and Justin Wolfers' scrutiny of studies purporting to show deterrent

---

[149] An example of a "null hypothesis" would be that in the true relationship being estimated by a regression equation, the coefficient for one of the independent variables in fact is zero, indicating that, after controlling for other variables, there is no relationship between the dependent variable and that independent variable.

[150] Some researchers have sought to counter this by encouraging the publication of statistically insignificant results. *See, e.g.*, Huai Yong Cheng, *The Potential Value of Negative Studies*, 6 J. AM. MED. DIRECTORS ASS'N 426 (2005).

[151] *See* J. Bradford De Long & Kevin Lang, *Are All Economic Hypotheses False?*, 100 J. POL. ECON. 1257 (1992) (conducting a statistical analysis of the distribution of statistical results to estimate the proportion of unrejected null nypotheses that are false). The De Long and Lang statistical analysis rejects "the null hypothesis that more than about one-third of *unrejected* null hypotheses … are true." *Id.* at 1258. That is, among published findings that do *not* show statistically significant outcomes

[152] *See, e.g.*, WILLIAM H. GREENE, ECONOMETRIC ANALYSIS 316-50 (3d ed. 1998) (providing an introduction to these issues).

[153] *See id.* (considering the possibility of nonlinear specifications).

[154] There may be flexibility both in determining the general coverage of the study (for example, what years or states to study), as well as in identifying outliers. Typically, when an observation is identified as an outlier, a researcher will run a regression both with and without the outliers to determine whether the results are robust. There are also econometric techniques designed to produce estimates not overly susceptible to outliers. *See, e.g.*, PETER J. ROUSSEEUW & ANNICK M. LEROY (2003). Some researchers, however, may not use these techniques.

[155] T. Dudley Wallace, *Pretest Estimation in Regression: A Survey*, 59 AM. J. AGRIC. ECON. 431, 431 (1977) ("Given the nature of economic data, pretesting has been and probably will continue to be widely used in spite of sharp criticism.").

[156] The traditional *t* statistic will be inaccurate when researchers test numerous regression specifications and then focus only on those whose *t* statistics appear to produce statistically significant results. *See, e.g.*, Dmitry Danilov & Jan R. Magnus, *Forecast Accuracy with Pretesting with an Application to the Stock Market*, 23 J. FORECASTING 251 (2004).

[157] *See* Jeff Strnad, *Should Legal Empiricists Go Bayesian?*, 9 AM. L. & ECON. REV. 195, 197 (2007) (noting that in law, "the researcher is dealing with observational data that cannot be extended by additional experimentation").

effects of the death penalty.[158] For example, they criticized a study by Hashem Dezhbakhsh and Joanna Shepherd,[159] focusing first on a cross-sectional analysis of homicide trends in states that either abolished or adopted the death penalty. Dohonue and Wolfers argue that the same general trends existed in states that had not changed death penalty policy, and reanalyzed the data with a difference-in-differences approach. This produced statistically insignificant results.[160] Similarly, Donohue and Wolfers modify a regression performed by Dezhbakhsh and Shepherd by including year fixed effect rather than decade fixed effect variables,[161] and making the critical coefficient statistically insignificant. Donohue and Wolfers take similar approaches in questioning several other papers' results.[162] Their point is not that prior commentators necessarily made poor choices, but that the particular reported results are not adequately supported.[163]

Often, there will be some subjectivity involved in determining whether a study's results are sufficiently robust to justify causal inferences. This does not mean that every empirical question will yield an uncertain answer. But the death penalty is hardly the only debate about which scholarly experts have hotly contested empirical outcomes. Other recent examples in the criminological context include debates about whether abortion legalization is responsible for the decrease in the crime rate,[164] and whether statutes allowing citizens to carry concealed handguns lower violent crime rates.[165] Whatever the merits, academics and policymakers have not reached consensus on these questions. Even if the median voter or median decisionmaker would be swayed by empirical results, it will not be easy to determine what to believe.

---

[158] *See* John J. Donohue & Justin Wolfers, *Uses and Abuses of Empirical Evidence in the Death Penalty Debate*, 58 STAN. L. REV. 791 (2005).

[159] *See* Hashem Dezhbakhsh & Joanna M. Shepherd, *The Deterrent Effect of Capital Punishment: Evidence from a Judicial Experiment* (Emory Law & Econ. Research Paper No. 04-04, 2003), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=432621.

[160] Donohue & Wolfers, *supra* note 158, at 800-04.

[161] A year fixed effect variable is simply a binary variable corresponding to each year (except one), with a 1 indicating that an observation was from that year and a 0 indicating that an observation was not from that year. *See, e.g.*, PETER KENNEDY, A GUIDE TO ECONOMETRICS 311-15 (5th ed. 2003) (providing an introduction to fixed effects variables). Decade fixed effects do the same for decades.

[162] *See, e.g.*, Donohue & Wolfers, *supra* note 158, at 811 (showing that the findings of Lawrence Katz et al., *Prison Conditions, Capital Punishment, and Deterrence*, 5 AM. L. & ECON. REV. 318 (2003), were sensitive to the choice of an applicable time period); *id.* at 816-18 (finding that H. Naci Mocan & R. Kaj Gittings, *Getting Off Death Row: Commuted Sentences and the Deterrent Effect of Capital Punishment*, 46 J.L. & ECON. 453 (2003), results were sensitive to the choice of the length of a time lag).

[163] *See id.* at 836 (stating that "we do not believe that economic or econometric theory are sufficiently well developed" to justify "a particularly strong prior belief about the 'correct specification'").

[164] The paper that started the debate is John J. Donohue III & Steven D. Levitt, *The Impact of Legalized Abortion on Crime*, 116 Q.J. ECON. 379 (2001).

[165] At the center of this debate is the book JOHN R. LOTT, JR., MORE GUNS, LESS CRIME (1998).

Publication bias is a danger in randomized studies too.[166] But there is less room for identifying alternative empirical specifications given the centrality of the random variable. As Esther Duflo has noted, in retrospective studies, "[e]x post the researchers or evaluators define their own comparison group, and thus may be able to pick a variety of plausible comparison groups,"[167] but in a randomized study, the treatment and comparison groups will generally be clearly defined. There is still some danger that researchers will decide not to publish, but that danger is considerably reduced when governmental institutions have sponsored the research by supporting the randomization of policy and a particular set of researchers has promised to analyze the effects of the experiment. Indeed, governments can virtually eliminate the risk by requiring publication of experimental results as a condition of funding.

2. Pseudorandom Experimentation

Concerns about retrospective analyses have led econometricians to use techniques that allow clearer identification of the effects of policy decisions. Although these techniques can help determine whether an association between a policy change and some other phenomenon is causal, there is still sufficient subjectivity in the execution of these studies that such analysis will only rarely leave statisticians, let alone policymakers, in unanimous agreement. Even with, or perhaps especially with, the most sophisticated tools, statistical analysis requires expert judgments about the quality of any particular study. Although there can also be considerable debate about the quality of randomized studies, randomization at least solves the central issues that techniques such as instrumental variable estimation and regression discontinuity design address.

a. Instrumental variables studies

Instrumental variables techniques substitute for an explanatory variable a prediction of that variable based on one or more additional variables, the instruments. Each instrument must be correlated with the explanatory variable, but not correlated with the independent variable except

---

[166] Selective publication of results has been most clearly demonstrated in the medical arena, though the studies do not assess whether selective publication is a result of self-censorship by authors (perhaps because they do not want to suggest that a drug was ineffective) or by journals. *See, e.g.*, Eric H. Turner, M.D. et al., *Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy*, 358 NEW ENG. J. MED. 252 (Jan. 17, 2008) (analyzing which reviews of antidepressant agents submitted to the FDA were subsequently published).

[167] Esther Duflo, *Scaling Up and Evaluation*, 2004 ANN. WORLD BANK CONF. ON DEVEL. ECON. 341, 353.

through the correlation with the explanatory variable.[168] Consider, for example, Steven Levitt's attempt to determine the effect of police presence on crime.[169] A regression simply using the change in the crime rate as a dependent variable and police presence as an independent variable would produce a biased coefficient, because the choice of how many police to hire is endogenous and may depend in part on expectations of whether crime is likely to increase. Levitt developed a regression predicting the change in the number of police officers based on factors including whether a mayoral election was scheduled,[170] and then substituted the predicted values from this regression for the variable directly measuring police presence. Using the predicted change in police presence as an explanatory variable rather than the actual change in police presence allows the experimenters to isolate the effect of increased police presence attributable to what amounts to a random factor, the election year calendar.

Instrumental variables studies will not always be adequate substitutes for truly randomized studies, however.[171] There may be some subjectivity in the choice of instruments.[172] Donohue and Wolfers, for example, criticized death penalty studies that used instrumental variables techniques.[173] They argue that in the Dezhbakhsh et al. study, the instruments for executions affect the independent variable (homicide rates) directly. Donohue and Wolfers establish this in part by showing that the relationship exists even in the states that have never had the death penalty,[174] and also by showing that the variables are correlated with rates of crimes for which the death penalty does not even apply.[175] Although there are statistical tests that can be

---

[168] That is, the instrument must not be correlated with the error term in the regression equation using the original explanatory variable. *See generally* Joshua D. Angrist & Alan B. Krueger, *Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments*, 15 J. ECON. PERSP. 69 (2001).

[169] Steven D. Levitt, *Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime*, 87 AM. ECON. REV. 270 (1997).

[170] *Id.* at 277.

[171] The Levitt study encountered an unusual criticism, that Levitt had made a programming error that made his estimates seem more precise than they in fact were. Justin McCrary, *Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime: Comment*, 92 AM. ECON. REV. 1236 (2002) (pointing out the error in Levitt's calculations).

[172] In the context of estimating the effect of increased police presence on crime rates, later researchers have suggested other instruments. *See, e.g.*, Jonathan Klick & Alexander Tabarrok, *Using Terror Alert Levels to Estimate the Effect of Police on Crime*, 48 J.L. & ECON. 267 (2005) (using terror alert levels in Washington, D.C., as an instrument, because the city deployed more police when the alert levels were raised, even though the terror alert levels would not be correlated to ordinary crime rates). If multiple studies produce consistently similar results, then researchers may conclude that the results are robust. But there will be no objective answer to the question of whether results are sufficiently robust to justify policy reliance on them.

[173] Donohue & Wolfers, *supra* note 158, at 827.

[174] *Id.* at 827-28.

[175] *Id.* at 830.

used to assess the validity of instruments,[176] the authors of the original studies can still claim that the instruments used were the best available.[177] Casual empiricism about empiricism suggests that the persuasiveness of many instrumental variables studies will often be debated.[178] The increased complexity of an instrumental variables study provides an additional layer of complexity about which statisticians can argue. This additional layer increases the number of choices that researchers face, potentially increasing the danger that studies will be subject to publication bias.[179]

b. Regression discontinuity studies

Another approach to pseudorandom experimentation, regression discontinuity design,[180] may help establish causal relationships, but still demands subjective interpretive judgments. This design takes advantage of discontinuities in decisionmaking about individual cases. For example, M. Keith Chen and Jesse Shapiro note that federal prisoners are assigned to prisons based in part on a score reflecting the inmates' need for supervision.[181] Inmates with scores around certain thresholds are likely to be quite similar to one another, but are assigned different treatments based on which side of the thresholds they fall. The discontinuity serves as a natural experiment, and Chen and Shapiro use the experiment to assess whether harsher prison conditions reduce recidivism.[182]

Although useful,[183] such a natural experiment requires careful statistical analysis to ensure that the populations on either side of the line are sufficiently equivalent in relevant

---

[176] *See* Jerry A. Hausman, *Specification Tests in Econometrics*, 46 ECONOMETRICA 1251 (1978).

[177] *See, e.g.*, Paul H. Rubin, *Reply to Donohue and Wolfers on the Death Penalty and Deterrence*, ECONOMIST'S VOICE, Apr. 2008, *available at* http://bpp.wharton.upenn.edu/jwolfers/Press/DeathPenalty(Rubin).pdf, at 2 (noting that while the instrumental variables may have been correlated with crime rights, "previous researchers believed (often based on empirical testing) that the instruments were as uncorrelated with crime rates as one was likely to find").

[178] For another example, see Badi H. Baltagi & James M. Griffin, *The Econometrics of Rational Addiction: The Case of Cigarettes*, 19 J. BUS. & ECON. STAT. 449 (2001) (challenging an earlier paper using instrumental variables techniques, Gary S. Becker et al., *An Empirical Analysis of Cigarette Addiction*, 84 AM. ECON. REV. 396 (1994), and offering an improved model).

[179] Esther Duflo has argued that publication bias may also be more problematic for instrumental variables studies for a technical reason. *See, e.g.*, Duflo, *supra* note 167, at 353 ("[P]ublication bias may actually more than compensate for the reduction in bias caused by the use of an instrument, because they tend to have larger standard errors, and researchers looking for significant results will select only large estimates."

[180] For an overview, see Jinyong Hahn et al., *Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design*, 69 ECONOMETRICA 201 (2001).

[181] M. Keith Chen & Jesse M. Shapiro, *Do Harsher Prison Conditions Reduce Recidivism? A Discontinuity-Based Approach*, 9 AM. L. & ECON. REV. 1 (2007).

[182] They conclude that such conditions do not reduce, and may even encourage, recidivism. *See id.* at 20.

[183] One study performed a regression discontinuity analysis in a context in which a randomized experiment was also performed,

respects. For example, Chen and Shapiro compared the demographics of two similar populations, and found significant demographic differences.[184] Whether the populations are nonetheless similar in relevant ways requires some subjective judgment. Moreover, their analysis explicitly modeled the function "that relates an inmate's score to his probability of recidivism,"[185] including binary controls for the cutoffs.[186] A danger is that if there is an omitted variable in the analysis or a misspecification of the functional form, the coefficient estimates could be biased.

Regression discontinuity studies require sufficient analytical judgment that their improved statistical power may not translate to a greater likelihood that the findings will be accepted in the public policy process. For example, a paper by Saurabh Bhargava and Vikram Pathania takes advantage of the discontinuity in cellular telephone rates around 9 p.m.[187] Call volumes today increase discontinuously around the 9 p.m. threshold, but there has been no increase in accidents immediately after 9 p.m. relative to the period before cell companies began to offer free calling after 9 p.m. Nonetheless, policymakers have continued to proclaim cell phone driving as dangerous as driving under the influence of alcohol.[188] A subsequent researcher, meanwhile, has minimized the Bhargava and Pathania findings, pointing out that the effects of cell phone use around 9 p.m. could be different from the effects at other times.[189] Regression discontinuity studies do not necessarily have greater impact on the policy process than other studies, even for the relatively small number of issues for which they are feasible.

---

finding that the results were similar. *See* Hielke Buddelmeyer & Emmanuel Skoufias, *An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA* (World Bank Policy Research Working Paper No. 3386, 2004).

[184] *See* Chen & Shapiro, *supra* note 181, at 16.

[185] *Id.* at 17.

[186] *Id.* at 20.

[187] Saurabh Bhargava & Vikram Pathania, *Driving Under the (Cellular) Influence: The Link Between Cell Phone Use and Vehicle Crashes* (AEI Working Paper No. 07-15, July 2007), *at* http://aei-brookings.org/admin/authorpdfs/redirect-safely.php?fname=../pdffiles/WP07-15_topost.pdf.

[188] *See, e.g.*, Mike Stuckey, *Hands-Free Phones Are Lifesavers, Study Says*, MSNBC, May 13, 2008, at http://www.msnbc.msn.com/id/24580099/ (quoting a California legislator who embraced a recent study on cell phones while apparently paying no heed to the Bhargava and Pathania study).

[189] *See* Jed Kolko, *The Effects of Mobile Phones and Hands-Free Laws on Traffic Fatalities* 5 (Pub. Pol'y Inst. of Cal. Working Paper No. 2007/07) (noting that Bhargava and Pathania acknowledged this possibility). Kolko's study includes a difference-in-differences analysis of the impact of hands-free laws that shows a statistically significant effect of the laws at the 0.05 level only for fatalities in wet road conditions. *Id.* at 32 tbl.7. A separate specification produced results at a 0.01 statistical significance level when a hands-free law was in effect 25 months or more, but this applied only to a single state. *Id.* at 35 tbl.10. Omitted variable bias could explain some of Kolko's results. For example, Kolko notes that there were "contemporaneous policy changes" related to alcohol consumption by drivers and that this eliminates the statistical significance at the 0.05 level for some results. *Id.* at 20.

3. The Laboratory of the States Reconsidered

For statistical research to influence policy, rather than merely serve as a sound bite, it must not only be executed well, but also be executed in a way that policymakers can understand and cannot ignore. These challenges pose hurdles for a frequent justification of federalism, that allowing states to make independent choices provides for a kind of laboratory testing of policy.[190] Susan Rose-Ackerman has shown that federalism may insufficiently promote experimentation for numerous reasons,[191] for example because one state may hope to free-ride on the activities of other governments.[192] Edward Rubin and Malcolm Feeley have similarly noted that experimentation sometimes may be expensive and likely not on balance beneficial for the experimenter, [193] and so centralized coordination may be needed to take full advantage of federalism. [194]

Yet, at least sometimes, states do change their policies and take risks in doing so in the hope of achieving informational benefits.[195] As Barry Friedman notes, states may innovate for a variety of reasons, quite apart from any desire to engage in "experimentation."[196] These state innovations serve at least a crude experimentation function.[197] Commentators may observe that one state's approach to a particular issue, such as health care reform,[198] has gone particularly

---

[190] The classic statement of this theory is Justice Brandeis's. *See* New State Ice Co. v. Liebmann, 285 U.S. 262, 311 (1932) (Brandeis, J., dissenting) ("It is one of the happy incidents of the federal system that a single courageous state may, if its citizens choose, serve as a laboratory; and try novel social and economic experiments without risk to the rest of the country."). For a discussion of this justification for federalism, Ann Althouse, *Vanguard States, Laggard States: Federalism and Constitutional Rights*, 152 U. PA. L. REV. 1745, 1750-76 (2004).

[191] *See* Susan Rose-Ackerman, *Risk Taking and Reelection: Does Federalism Promote Innovation?*, 9 J. LEGAL STUD. 593 (1980).

[192] *Id.* at 594. The possibility that there might be insufficient incentives to innovate is apparent even in areas in which state competition has generally been trumpted, such as corporate governance law. *See* Michael Abramowicz, *Speeding up the Crawl to the Top*, 20 YALE J. ON REG. 139 (2003) (arguing that there are suboptimal incentives for states to innovate in corporate law). *But see* Roberta Romano, *The States as a Laboratory: Legal Innovation and State Competition for Corporate Charters*, 23 YALE J. ON REG. 209 (2006) (arguing that states are in fact effective laboratories in the corporate charter context).

[193] Edward L. Rubin & Malcolm Feeley, *Federalism: Some Notes on a National Neurosis*, 41 UCLA L. REV. 903, 923-26 (1994).

[194] *Id.* at 926 (noting that absent coordination by a central authority, "state-initiated experiments are unlikely to be truly useful to other states because of more specific, technical variations" among the states).

[195] *See, e.g.*, FERC v. Mississippi, 456 U.S. 742, 787-88 (1982) (O'Connor, J., concurring in part and dissenting in part) ("[S]tate experimentation is no judicial myth.").

[196] Barry Friedman, *Valuing Federalism*, 82 MINN. L. REV. 317, 399 (1997) ("'Innovation' might have been a better word choice for Justice Brandeis than 'experimentation,' saving us all a lot of bother.").

[197] Dorf and Sabel express more confidence in the ability of state innovations to improve knowledge, as long as there is some centralized evaluation of state activities. *See* Dorf & Sabel, *supra* note 1, at 345 (explaining how administrative agencies can serve as "the continuing organized link between the national and the local, helping to create through national action the local conditions for experimentation, and changing national arrangements accordingly.").

[198] *See, e.g.*, Sara Rosenbaum, *Mothers and Children Last: The Oregon Medicaid Experiment*, 18 AM. J.L. & MED. 97 (1992).

badly or well, and this may influence their decisionmaking. Federalism, however, does not easily facilitate a scientific approach to experimentation. The difficulty that social scientists and especially policymakers face in assessing the results of state innovations contributes to the inaptness of the states-as-laboratories metaphor.

Still, federalism may be more conducive to experimentation than alternatives. Previous commentators have noted that randomized experiments are much more common in North America than in the rest of the world,[199] and speculated that federalism may help explain this.[200] In any event, the mere existence of different jurisdictions could be conducive to randomized experimentation in two ways. First, it may be possible to randomize policies across states, at least among states that consent. It would be more awkward to randomize policies in the absence of generally accepted jurisdictional boundaries. And second, states themselves can serve as loci for experimentation at smaller jurisdictional levels, such as cities and counties. Indeed, randomized experiments have increasingly been conducted within states.[201] Perhaps federalism will someday produce a stream of experimental data. The next section will argue, however, that even this may be insufficient absent self-execution.

B. Limits of Randomization Studies

Advocates of randomized studies have emphasized that only this type of study can establish causality with high confidence. For example, Esther Duflo has argued that "while it is always possible to reject experimental results on the grounds that the experiment was poorly designed, or failed, when an experiment is correctly implemented (which is relatively easy to ascertain), there is no doubt that it gives us the effect of the manipulation that was implemented."[202] But what "is relatively easy to ascertain" may still remain controversial in public debate. Moreover, even if the measured effects can be confidently traced to the "manipulation," some extrapolation will generally be needed to assess the full consequences of a

---

[199] *See* GREENBERG ET AL., *supra* note 3, at 38 (noting as an exception that the Netherlands tested an unemployment-counseling program).

[200] *Id.* One justification for this is that "[f]ederal funds for particular programs may be used with considerable discretion by the states, and this fact has encouraged the view that the states should literally be the laboratories of democracy." *Id.*

[201] *See* GREENBERG ET AL., *supra* note 3, at 37-38.

[202] Esther Duflo, *Field Experiments in Development Economics* (Jan. 2006), http://econ-www.mit.edu/files/800, at 23.

law enacting the legal experiment. This section explains that this is so because of difficulties with both measurement and replication.

The problems identified here can generally afflict nonrandomized studies as well. The argument for randomized rather than nonrandomized studies, though, cannot be only that randomized studies are more definitive. Even if policymakers and public debate recognize the benefits of randomization, there will still be enough room for debate that these studies' impact is not likely to be proportional to the quality of information that they produce. Randomized studies do, however, facilitate self-execution. The imperfections and likely effects of randomized studies can be gauged by policymakers relatively easily in advance of the studies. As a result, policymakers should be able to commit in advance to particular legal outcomes dependent on the studies' results. At least ex ante, the results of randomized experiments will generally have clear implications, if not for exactly what policy should be, then at least for whether there is greater justification than before for moving policy in a particular direction. That is not true for nonrandomized experiments, which will generally not produce a single outcome measure or measures that policymakers could easily agree on in advance.

1. Measurement Problems

a. Elusive effects

Legal policies can have easy-to-measure and hard-to-measure effects. Numerical assessments of randomized experiments will tend to focus on easy-to-measure variables, but observers may care about other variables too. Suppose, for example, that a jurisdiction is considering adopting criminal shame sanctions.[203] Even placing aside nonconsequentialist concerns, one might care about how much a shame sanction would reduce recidivism and deter crime. The former is likely easier to measure than the latter, and hardest of all to measure would be whether the shame sanctions degrades society, indirectly producing other adverse consequences.[204] If some observers care mostly about recidivism, while others care mostly about general deterrence and social degradation, experimentation might not be of much use. Such a

[203] *See generally* Stephen P. Garvey, *Can Shaming Punishments Educate?*, 65 U. CHI. L. REV. 733 (1998) (arguing that shame sanctions may promote moral educate); David R. Karp, *The New Debate About Shame in Criminal Justice: An Interactionist Account*, 21 JUST. SYS. J. 301 (2001) (arguing in favor of some and against some shame sanctions).

[204] *See, e.g.*, James Q. Whitman, *What Is Wrong with Inflicting Shame Sanctions?*, 107 YALE L.J. 1055, 1059 (1998) (arguing against shame sanctions on the ground that they involve "a specifies of official lynch justice").

dispute concerns values rather than empirics. If, however, recidivism is important to both proponents and opponents of shame sanctions, then experimentation may be useful, even if the results of such an experiment would not produce consensus. As long as there is disagreement ex ante about the shame sanction's effect on recidivism, with opponents of the sanction more likely to doubt that it would reduce recidivism, parties who could not agree on the merits might agree to a self-executing randomized experiment.[205]

Randomized experimentation, however, may fail to bring parties closer together if policymakers on one side of an issue believe that achieving success on an easy-to-measure variable would be correlated with failure on hard-to-measure variables. Consider, for example, a law that implements standards-based education reform, such as the No Child Left Behind Act.[206] Some worry that the Act will push schools to focus only on measurable proxies for student learning, such as multiple choice test scores.[207] The experimentation literature labels this the "multitasking problem,"[208] where "individuals facing high-powered incentive schemes may change their behavior in such a way that the proximate outcome on which the rewards are based increases, but the ultimate outcome in which the principal is interested remains constant or even decreases."[209] A self-executing experiment could test standards-based reforms, with different rules applied to randomly selected states or school districts. But if opponents believe that these reforms improve multiple choice scores at learning's expense, they will resist a self-executing experiment dependent on such scores, unless they believe that proponents have exaggerated even the reforms' effects on those scores. Any agreement on a self-executing experiment would depend on development of some alternative measurement metric. For example, a randomly selected subset of students might sit for an exam with essay and open-ended questions, under an agreement that the grades would be used to assess the reforms but not to determine district or teacher outcomes.[210]

---

[205] *See supra* Part II.A.2.

[206] Pub. L. No. 107-110, Jan. 8, 2002, 115 Stat. 1425.

[207] *See, e.g.*, Lisa Kelly, *Yearning for Lake Wobegon: The Quest for the Best Test at the Expense of the Best Education*, 7 S. CAL. INTERDISC. L.J. 41 (1998).

[208] *See* Bengt Holmstrom & Paul Milgrom, *Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design*, 7 J.L. ECON. & ORG. 24 (1991).

[209] *See* Duflo, *supra* note 202, at 5.

[210] There already exists a federal test, known as the National Assessment of Educational Progress, that is applied to a subset of students and not used to assess individual students or schools. *See* James E. Ryan, *The Perverse Incentives of the No Child Left Behind Act*, 79 N.Y.U. L. REV. 932, 943-44 (2004). But the test is not currently used to test the Act itself (with some school

At times, policymakers might commit to resolving a randomized experiment based on a cost-benefit analysis, conducted pursuant to the professional standards generally applicable to such analyses.[211] Indeed, cost-benefit analysis has increasingly been used in measuring the results of randomized experiments,[212] and a virtue of cost-benefit analysis is that it reduces diverse variables to a single metric, dollars.[213] Critics of cost-benefit analysis, however, complain that such analysis sometimes inappropriately makes value choices, in particular making hard variables more important than soft variables.[214] Cost-benefit analysis can serve as a viable procedure for measuring experimental outcomes where both sides believe that cost is a reasonable proxy for the effectiveness of an experiment, or where the threshold for success is adjusted based on the perception that the analysis might shortchange certain variables that would favor one position.

b. Imperfect randomization

A computer can randomize between treatment and control groups, but it is not always straightforward to ensure that the treatment group receives the treatment and the control group does not. Dangers include attrition (where some randomized individuals or entities drop out of a study), crossover (where some control group members receive the treatment, or vice versa), and spillovers (where the treatment has some effect on the control group as well). None of these problems, however, presents an insurmountable challenge to self-executing randomized experiments.

i. Attrition

Attrition is a problem not merely because it decreases the size of the sample, but also because it may bias experimental results when the attrition rate depends on selection for

---

districts assigned to the Act's regime and some assigned to another regime).

[211] *See, e.g.*, Exec. Order No. 12,866, 58 Fed. Reg. 51,735 (Sept. 30, 1993).

[212] Increasingly, cost-benefit analysis has been used to test programs. GREENBERG ET AL., *supra* note 3, at 36-37 (noting that only 27% of early social experiments used cost-impact comparisons, but that 60% of more recent tests did, and that recently there "was an attempt to measure as many benefits and costs in monetary terms as possible and then to compare the benefits received by all members of society to the costs incurred by all members of society").

[213] Experiments may have even greater political saliency when they demonstrate that they produce some benefit at no government cost. The Unemployment Bonus Experiment results, for example, indicated that in many treatments, there was net savings to the federal government. That is, paying bonuses for leaving unemployment led to less governmental expenditure. *See, e.g.*, GREENBERG ET AL., *supra* note 3, at 184. Ayres describes how this actually affected the government's decision on the program. *See* AYRES, *supra* note 62, at 65-67.

[214] *See, e.g.*, Lisa Heinzerling, *Regulatory Costs of Mythic Proportions*, 107 YALE L.J. 1981, 2042-69 (1998) (elaborating on this argument).

treatment. Consider, for example, studies assessing improvements made by schools in a developing country. A school's randomization into a comparison group might increase drop-out rates.[215] If the drop-outs tend to be the weaker students, and if the measurement of school success depends on tests of current students, then the attrition produces an artificial hurdle for the treatment. Attrition can also bias results when randomization occurs at the level of the individual. In a medical study, for example, people who receive the treatment but then suffer severe side effects might refuse to participate further in the study, making those who continue with the treatment a nonrepresentative sample.

Given any degree of attrition, those reviewing a study may argue about the best interpretation of the results. Statisticians may attempt to impute measurements for those who drop out, by comparing their characteristics with those of other subjects.[216] But this solution is imperfect, because there might be some unmeasurable difference between those who continue in an experiment and those who drop out. Ultimately, sound statistical judgment is needed to assess such models' reliability. In a public policy context, those who do not like the normative implications of a randomized experiment will often be able to criticize it, rightly or wrongly, based on attrition.

A more administrable solution is to use matched samples.[217] If someone from the treatment group drops out, results of the corresponding match from the control group are not counted either. This approach can be used also when randomization is at the institutional or jurisdictional level, if individuals can be matched across institutions or jurisdictions. With matching, it is not necessary ex post to construct a model that seeks to correct for attrition bias, which would increase the danger of subjectivity or manipulation. Statisticians would be needed to assign the original matches based on observable characteristics, but the matching would be difficult to manipulate, before it is known who will drop out.

---

[215] *See, e.g.*, Abhijit V. Banerjee et al., *Remedying Education: Evidence from Two Randomized Experiments in India*, 122 Q.J. ECON. 1235, 1245 (2007) (discussing this problem).

[216] *See, e.g.*, Richard B. Miller & David W. Wright, *Detecting and Correcting Attrition Bias in Longitudinal Family Research*, 57 J. MARRIAGE & FAMILY 921, 923 (1995) (describing the standard method of incorporating a variable representing the probability of dropping out directly into the stud

[217] *See* ESTHER DUFLO ET AL., USING RANDOMIZATION IN DEVELOPMENT ECONOMICS RESEARCH: A TOOLKIT 35-36 (2006), http://www.povertyactionlab.com/papers/Using%20Randomization%20in%20Development%20Economics.pdf ("An extreme version of blocked design is the pairwise matched design where pairs of units are constituted, and in each pair, one unit is randomly assigned to the treatment and one unit is randomly assigned to the control.").

This approach has a significant disadvantage. It throws away data that a neutral statistician might find valuable. Once again, a simplified statistical procedure may be superior in a public policy context to a more sophisticated procedure that a neutral statistician might choose, because the simple procedure is less manipulable. The simplified procedure does not eliminate the problem of attrition bias. Unmeasurable differences may lead to imperfect matching and some residual attrition bias. This should not pose an insurmountable problem for ex ante commitment or negotiation. Either the direction of the residual bias is unpredictable, in which case it should not thwart negotiations ex ante, or policymakers will anticipate some degree of bias, in which case this may affect the thresholds at which self-executing policies will take effect.

ii. Crossover

Legal experimentation may be less vulnerable to crossover than other social experimentation. When a particular legal regime is assigned to some individual or entity, that is not easy to escape. But imperfections may occur nonetheless, especially if the government wishes to leave some room for later discretion. For example, a social experiment on police response to domestic violence randomized those accused of domestic violence so that only some were chosen for automatic arrest.[218] Officers, however, were given the discretion to arrest those not randomly selected, and so some individuals in the control group migrated to the treatment group.[219] Crossover can also occur if well-connected people can thwart random assignment. Alan Krueger, studying the effect of student to teacher ratios, found that some parents had managed to convince schools to reallocate their children from large to small classes.[220] This dilutes the treatment, as the small classes become larger, and means that the treated students on average will come from relatively highly motivated families.

Once again, statistical correctives exist. Under an "intention-to-treat" methodology,[221] an individual or entity who crosses over is counted with the group to which that person or entity

---

[218] *See* Lawrence W. Sherman & Richard A. Berk, *The Specific Deterrent Effects of Arrest for Domestic Assault*, 49 AM. SOCIOLOGY REV. 261 (1984) (finding that arrest of domestic violence suspects deterred subsequent violent behavior).

[219] *See id.* at 264 (discussing the attrition problem and noting that 78% of perpetrators assigned to receive advice were arrested instead).

[220] Alan B. Krueger, *Experimental Estimates of Education Production Functions*, 114 Q.J. ECON. 497, 505 (1999) (reporting higher attrition rates of students in smaller classes).

[221] *See, e.g.*, Guido Imbens & Joshua Angrist, *Identification and Estimation of Local Average Treatment Effects*, 62 ECONOMETRICA 467 (1994) (discussing this approach)..

was originally assigned. This reduces the measured effect of the treatment. Statisticians can compensate for the bias introduced by the intention-to-treat approach with a simple mathematical formula.[222] It is a close call whether it makes sense in a randomized legal experiment for this adjustment to be specified ex ante. This compensation makes the experiment more obtuse. Assuming, however, that it is possible to measure who ended up receiving the treatment and who ended up receiving the control, the formula can be applied mechanically, without allowing any discretion to the experimenters, and will generally improve the estimate of the treatment effect.

This is an imperfect adjustment, because those who cross over may differ systematically from those who do not. In theory, more elaborate adjustments are possible, with models projecting what the outcomes of those who cross-over would have been had they not. But once again, although many statisticians would think this sensible, for public policy purposes, any improvement in accuracy may not be worth the increase in subjectivity and manipulability. If crossover bias is an acute concern, policymakers can insist on the same solution as suggested above for attrition bias, using matched samples and excluding both the observations that cross over and their matches. This results in throwing away twice the data, and so may bring us even further from what statisticians might suggest, and it remains an imperfect solution given the possibility of imperfect matching of samples. But it may be enough to allow negotiations over self-executing experiments.

iii. Spillovers

The final danger, and at times the most troubling one for randomized legal experimentation, is that the treatment will spill over on the control group. Suppose, for example, that a shame sanction reduces recidivism not only in those who are shamed, but also in those who are randomized to the control group but hear about the shaming. Or, suppose that firms randomized to a relaxed securities disclosure regime decide that they want to disclose as much as their competitors. The comparison of treatment and control groups will underestimate the effects of the intervention. On the other hand, suppose that a random experiment eliminates speed limits on a random set of roads. Some drivers on the control roads may conclude that police, needing to

[222] As Esther Duflo explains, a statistician can "scale up the difference [between the treatment and the control group] by dividing it by the difference in the probability of receiving the treatment in those two groups." Duflo, *supra* note 167, at 354.

fill their time somehow, will devote extra attention to the control roads. If these drivers slow down, measurements of the speed differential will be exaggerated.

A sometimes feasible solution is to randomize across geographical areas rather than individuals. Edward Miguel and Michael Kremer showed that randomized studies at an individual level had underestimated the benefits of deworming drugs, which benefited those in the immediate area who had not received the drugs.[223] Randomizing across geographical areas largely solved the problem. This solution is not without drawbacks, however. Especially if the number of jurisdictions is small, a comparison of changes in treatment and control jurisdictions may not have much statistical power. In addition, some people may move to take advantage of the law elsewhere.[224] For example, if a study eliminates the minimum wage in randomly selected counties, then some workers may cross county lines in search of minimum wage jobs. This could have various effects on measurement, such as resulting in an underestimation of the increase in unemployment, if any, attributable to a minimum wage.[225]

Any attempts to correct statistically for the problem of spillovers will involve guesswork. Often no objective metric of whether a spillover has occurred exists. If the spillover is measurable, statisticians still need to devise corrective regression models. The best policy reaction then may be to ignore the spillover. This need not be problematic, so long as policymakers recognize what such an experiment measures. Many laws have both direct effects on regulated individuals or entities and indirect effects, typically via social norms.[226] A randomized experiment generally measures the extent of the direct effects over and above the indirect effects. With spillovers, what is measurable will therefore be a smaller part of the overall policy calculus. The lower the relevance of what is empirically measurable, the less beneficial randomized experimentation will be, but it can still be useful in facilitating policymaker agreement.

---

[223] Edward Miguel & Michael Kremer, *Worms: Identifying Impacts on Education and Health in the Presdence of Treatment Externalities*, 72 ECONOMETRICA 159 (2004).

[224] Randomization across geographic areas can produce Tiebout sorting in much the same way as endogenous policy selection. *See generally* Charles Tiebout, *A Pure Theory of Local Expenditures*, 64 J. POL. ECON. 416 (1956) (providing the seminal account of the effects of citizen mobility).

[225] These complexities may, however, be considerably smaller than those inherent in analyzing existing policies. *See, e.g.*, David Card & Alan B. Krueger, *Time-Series Minimum-Wage Studies: A Meta-Analysis*, 85 AM. ECON. REV. (PAPERS & PROCEEDINGS) 238 (1995) (reviewing various studies).

[226] *See, e.g.*, GREENBERG ET AL., *supra* note 3, at 15 ("[A]n important aspect of some policy innovations intended for widespread adaptation is that they cause changes in community attitudes and norms; these, in turn, result in feedback effects that influence the innovation's success.").

2. Generalizability

The generalizability concern is that the treatment group might behave differently if a policy were enacted permanently after the experiment, and thus, as with spillovers, the concern is that measured effects may differ from actual effects.

a. Awareness of experimental context

People in an experiment may behave or respond differently if they know they are in an experiment and if they know that they are in the treatment group. In medical experiments, a blind design can eliminate the latter problem but not the former.[227] With randomized legal experiments, it will generally not be possible to conceal from participants whether they have been assigned to a treatment or a control group. This makes possible both "Hawthorne effects,"[228] in which the treatment group responds to being treated, for example by working to ensure the experiment's success or failure, and "John Henry effects,"[229] in which the comparison group responds to not being treated.[230] Little can be done to prevent Hawthorne and John Henry effects or to adjust for them in other than a speculative way. These effects seem unlikely to be generally large, however, especially for policies that affect people in nontrivial ways. Ex post, any study result can be questioned on the theory that some Hawthorne or John Henry effect might have occurred, and so these effects furnish another reason that using experimentation simply to produce more information for the policy process is unlikely to be beneficial. But ex ante, policymakers can make armchair estimates of these effects and negotiate around them.

---

[227] The most obvious way that this might occur is through the placebo effect. Research has shown that patients will sometimes respond medically even when not treated if they believe that they might be. *See* Anup Malani, *Identifying Placebo Effects with Data from Clinical Trials*, 114 J. POL. ECON. 236 (2006) (verifying the existence of placebo effects on the basis of differential probabilities of random assignment to the treatment).

[228] *See, e.g.*, WILLIAM H. WHYTE, JR., THE ORGANIZATION MAN 34 (1956) (claiming that increased productivity in an industrial experiment by Hawthorne was a result of experimenters' behavior toward those treated). *But see* Stephen R.G. Jones, *Was There a Hawthorne Effect?*, 98 AM. J. SOCIOLOGY 451 (1992) (questioning the existence of the effect by scrutinizing Hawthorne's original study).

[229] *See, e.g.*, Allen C. Barrett & Doris A. White, *How John Henry Effects Confound the Measurement of Self-Esteem in Primary Prevention Programs for Drug Abuse in Middle Schools*, 36 J. ALCOHOL & DRUG EDUC. 87 (1991) (providing an alleged example of John Henry effects).

[230] DUFLO ET AL., *supra* note 217, at 68 ("The comparison group may feel offended to be a comparison group and react by also altering their behavior (for example, teachers in the comparison group for an evaluation may 'compete' with the treatment teachers or, on the contrary, decide to slack off).").

b. Differences in experimental context

Even absent behavior changing because of knowledge of the experiment, the experimental context may differ from the context in which a policy ultimately would be implemented. The experiment might affect a different population, be on a smaller scale, involve a different legal change, test only marginal policy changes, occur for only a limited period of time, or involve greater or lesser commitments of resources.

The population may differ if an experiment is tried in only one location or only with some nonrandom subset of the individuals and entities who would eventually be affected by a law. Cost considerations may justify such nonrepresentativeness, and indeed it is common for "demonstration projects" to be deployed in one or more particular regions rather than randomly.[231] Sometimes, it may be feasible to conduct randomization at a national level, for example in choosing Medicare recipients who will receive extra follow-up phone calls, but coordination and data-gathering needs may make this difficult. At times, skepticism about inferences from an experiment on a nonrepresentative population may be justified.[232] For example, a randomized securities disclosure experiment on a sample of small firms might not extrapolate easily to a sample of large firms.[233]

Scale may be an even more important concern. A common criticism of laboratory experiments is that people may not behave as they would in other decisionmaking contexts, because the stakes are too trivial.[234] Similar problems can affect randomized experiments. Suppose, for example, that the federal government tested the minimum wage by randomly selecting one percent of adults, allowing those selected the option of informing employers that they would not need to be paid minimum wage. In theory, eliminating the minimum wage might increase employment. But businesses may not think it worthwhile to change their hiring

---

[231] For a discussion of the transition from local demonstration projects to projects on a national scale, see Duflo, *supra* note 167, at 342-45.

[232] *See, e.g.*, GREENBERG ET AL., *supra* note 3, at 15 (""[I]mpact estimates frequently are limited to relatively few geographic areas. Because the sites are rarely selected randomly, the external validity of the evaluations can be questioned.").

[233] It is possible, for example, that the expense of disclosure is burdensome for small firms but trivial for large ones. *Cf., e.g.*, Paul Rose, *Balancing Public Market Benefits and Burdens for Smaller Companies Post Sarbanes-Oxley*, 41 WILLAMETTE L. REV. 707, 733-34 (2005) (considering evidence that the Sarbanes-Oxley Act is causing small firms especially to go private). Similarly, disclosure may be more important for large firms with many divisions and operations. On the other hand, it may be easier to execute frauds at smaller firms, making disclosure more important for these firms.

[234] Duflo, *supra* note 202, at 21. This helps explain why researchers studying social norms through ultimatum games have experimented in developing countries, where it is feasible to make the stakes large enough to affect experimental subjects' welfare. *See, e.g.*, Robert Slonim & Alvin E. Roth, *Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic*, 66 ECONOMETRICA 569 (1998).

practices, or to risk dissension from inconsistent wages, to have the chance to hire a few workers at a lower wage. Data from such a study therefore might not reliably reflect the effect of eliminating the minimum wage.

Randomized experiments also cannot be used to assess uncontroversially policies other than the ones tested. Suppose, for example, that a randomized test of laws allowing possession of concealed handguns were instituted, with different counties in a state randomly assigned to different regimes. The debate on the "more guns, less crime" hypothesis has largely focused on such statutes,[235] but the broader debate concerns many government policies affecting the incidence of handgun ownership. Those who believe that laws permitting concealed handguns reduce crime may think that this supports relaxation of other gun control rules (such as mandatory waiting periods), while those who believe that such laws increase crime may think that this supports broader gun control (such as handgun registration requirements). A randomized study presumably would shed some light on these issues, and perhaps legislators on both sides could agree to a bidirectional self-executing legal experiment of concealed carry laws that could result in collateral legal changes. Because other gun control policies could have different effects, it is unlikely that such a randomized experiment, if not self-executing, would have much indirect policy influence on these other questions.

Even where the same legal policy is at issue, the legal changes effected by an experiment will generally be temporary, and responses to temporary laws may be different from responses to permanent laws. Sometimes, this danger will be obvious. For example, tax theorists recognize that taxpayers may change the timing of income recognition to take advantage of anticipated changes in tax rules.[236] As a result, randomly granting some taxpayers a one-percent tax decrease in a particular year as a test of supply-side taxation theories would likely exaggerate the effects of a decrease in taxes on income.[237] Policymakers might design the study to more closely approximate the effects of a permanent law, perhaps by focusing only on components of income whose recognition is not as easily timed. But unless experiments are permanent for the affected

---

[235] *See, e.g.*, LOTT, *supra* note 165.

[236] *See, e.g.*, Daniel N. Shaviro, *An Efficiency Analysis of Realization and Recognition Rules Under the Federal Income Tax*, 48 TAX L. REV. 1 (1992) (providing an economic analysis of recognition).

[237] This point may apply just as strongly to experiments on government support. Commentators have pointed out that individuals may work more during an experiment providing them government services if they believe that the support will not continue through time. *See, e.g.*, Frank P. Stafford, *Income-Maintenance Policy and Work Effort: Learning from Experiments and Labor-Market Studies*, *in* SOCIAL EXPERIMENTATION, *supra* note 3, at 95, 101.

population (for example, through a commitment to give some taxpayers a permanent drop in tax rates), there will almost always be some subjective extrapolation needed to determine how much the temporary nature of an experiment affected outcomes.

Sometimes an experiment measures only marginal effects, either because the experiment is temporary or because the experiment explicitly limits itself to an intervention at the margin.[238] There is no guarantee that marginal effects will at least correctly identify the approximate impact of the policy. For example, in the hypothetical concealed carry experiment, a permanent law might encourage more people to possess concealed handguns than a temporary law, but it is not clear how the additional group of adopters differs from the group that responds even to the temporary law. Perhaps the initial responders will tend to include more criminals seeking to take advantage of the law and the subsequent group will be more law-abiding, but this is only speculation. The early and late adopter groups may differ culturally, and an exploration of these differences would require controversial analysis.

Sometimes, a temporary law may be a poor proxy for long term effects because the law will have dynamic as well as static effects. Studies, for example, that seek to assess private school choice plans may fail to capture what proponents of such plans claim is a principal benefit, that free enterprise will allow educational entrepreneurs to learn over time what works.[239] Other arguments, however, suggest that a static analysis might overestimate the benefits of free choice; for example, in the short term for-profit schools might be willing to lose money in the hope of increasing the chance of being permitted to continue to receive public funds in the future. As another example, critics of the time-of-use electricity experiments argued that with a longer term study, customers would buy appliances that would help them adjust their electricity use based on time of day.[240] Whether or not this argument is persuasive, the ease of constructing such an argument helps explain why past randomized experiments have not had greater public policy effects.

---

[238] *See, e.g.*, Dean Karlan & Jonathan Zinman, *Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts* (June 25, 2007), *available at* http://ipa.phpwebhosting.com/images_ipa/ExpandingCreditAccess.v3.pdf (reporting an experiment in which lending criteria were loosened for randomly selected marginal borrowers).

[239] *See, e.g.*, Terry M. Moe, *Beyond the Free Market: The Structure of School Choice*, 2008 BYU L. REV. 557, 571 (2008) (making this point).

[240] *See, e.g.*, JOSKOW at 46 (noting that "these experiments only allow us to estimate short-run elasticities of demand, given existing appliance stocks").

One or more of the above problems will affect, to a lesser or greater degree, most randomized experiments. But randomized experiments still provide useful information and should change a neutral Bayesian policy analysts' priors about the effects of policy interventions, sometimes dramatically. Often, though, the policy response to randomized experiments seems likely to be less than the informational content of the experiments would justify. Self-executing randomized experiments, even where improving information about only one facet of a complex policy problem, can combat this. There is a danger that advance commitment will lead to unjustified legal responses to experimental change, but for many of the problems detailed above, policymakers should be able to anticipate the limitations of the experiments. In a legal culture that embraced experimentation, policymakers could negotiate and commit to proportional legal changes in response to randomized experiments.

IV. GUIDELINES FOR AND OBJECTIONS TO LEGAL EXPERIMENTATION

A. Guidelines

In many respects, randomized, self-executing experiments should conform to ordinary principles of experimentation. For example, there should be a large enough sample to generate meaningful results.[241] There is no magic number for all experiments; a small number of observations may be enough if the measured effect of the intervention is anticipated to be large, but a large number may be needed for small anticipated measured effects. Those who hope that a self-executing experiment will produce results in their favor have some incentive to ensure that the sample is sufficiently large that noise is unlikely to obscure general effects. The higher the number of observations, the better chance that any actual effect will correctly be identified as existing at any particular threshold of statistical significance. Policymakers need not, however, choose any particular level of statistical significance, such as 0.05, as the threshold for self-execution. Statisticians have long recognized these thresholds as arbitrary.[242] Instead, policymakers might identify cutoffs for self-execution based on raw numbers, for example promising to adopt a new policy if the treatment group performs 10% better on some measure than the control group. This enhances the comprehensibility of the experiment.

---

[241] *See, e.g.*, DUFLO ET AL., *supra* note 217, at 29 (discussing the issue of sample size in randomized experiments).

[242] *See, e.g.*, Lester V. Manderscheid, *Significance Levels, 0.05, 0.01, or ?*, 47 J. FARM ECON. 1381 (1965) (urging that the applicable level of statistical significance be tailored to a particular purpose).

To set cutoffs, policymakers need some understanding of what the treatment and control policies are. Legislators need not specify all details; they may delegate these decisions to an administrative agency. Ideally, the treatment and control policies should produce meaningful information about laws that might be self-executed after the experiment. Meanwhile, policymakers must consider the unit of analysis at which randomization occurs.[243] If randomization is at the jurisdictional or institutional level, then even if there are many affected individuals or entities, the number of independent observations is the number of separately randomized units. Statistical analysis could be used to assess individual responses to policies, but only at the risk of reintroducing omitted variable bias. Finally, policymakers should generally use matched samples, with matching occurring before the experiment on all available variables, to reduce attrition bias.

A final, but more controversial, design suggestion is to avoid problems of self-selection and attrition by making participation mandatory. Social experiments to date have largely been opt-in, allowing individuals to choose whether to participate and then perhaps also whether to opt out.[244] This is not surprising given the conventional view of social experimentation as a form of academic research. Academics cannot experiment on research subjects without informed consent.[245] But governments in theory could make participation in a randomized experiment mandatory and even institute reporting requirements. There will always be some people who ignore the rules, and some unavoidable attrition, due to factors like emigration and death. But a government could either not count such individuals (and their matches) or develop some other convention for how to count them.[246] The question is whether experiments without informed consent would be ethical.

---

[243] *See, e.g.*, DUFLO ET AL., *supra* note 217, at 40.

[244] Alice M. Rivlin & P. Michael Timpane, *Introduction and Summary*, *in* ETHICAL AND LEGAL ISSUES OF SOCIAL EXPERIMENTATION, *supra* note 10, at 1, 7.

[245] *See generally* Kathryn A. Tuthill, *Human Experimentation*, 18 J. LEGAL MED. 221 (1997) (providing a legal overview).

[246] The convention might depend on context. For example, in an experiment on securities disclosure, the bankruptcy of a corporation could count as stock price declining to zero. An individual's death might count as a bad result in a health care policy experiment, but simply be ignored in an experiment on fee shifting in court cases.

B. Objections

1. Objections to Randomness

a. Ethical concerns

This Article's treatment of the ethics of randomized legal experiments will be brief for two reasons. First, the Article's general argument does not depend on a resolution of whether the government must obtain informed consent. Even with an informed consent requirement, randomized experimentation could still occur for many policies. For example, there will generally be no ethical objections to an experiment like the Medicare experiment,[247] where any participant may choose not to receive the services offered by the government. (There may be objections based on inequality among those who volunteer for the experiment, an issue to which the Article will return below.[248]) Second, an existing collection of essays already explores this issue in considerable detail.[249]

This section will briefly summarize and develop the argument that legal experimentation imposes no ethical hurdles beyond those inherent in general legal policymaking, while also sketching the opposing position. The argument against an informed consent requirement distinguishes legal from medical experimentation,[250] where informed consent is generally required.[251] An unconsented-to medical treatment violates a patient's bodily integrity rights.[252] The problem is not randomization. A state could not insist that all of its citizens take a new drug. Any rights that the individual has against the state constrain the state in legal experimentation. So, if a person has a right not to have property taken by the state,[253] then the state cannot take

---

[247] *See supra* note 13 and accompanying text.

[248] *See infra* Part IV.B.1.b.

[249] *See* ETHICAL AND LEGAL ISSUES OF SOCIAL EXPERIMENTATION, *supra* note 10.

[250] Rivlin and Timpane summarize this argument as follows:

> [S]chool officials make decisions all the time that involve adoption of new curricula or educational approaches without firm knowledge of what the effects will be. There is always some chance of harm to some or all children which has to be weighed against the possible benefits of the change. Calling the change an 'experiment' does not alter the moral dilemma involved or call for special rules. Such rules might have the perverse effect of putting special obstacles in the way of careful examination and evaluation of change, while allowing quite drastic changes that had no experimental or tentative flavor to proceed unquestioned.

Rivlin & Timpane, *supra* note 244, at 5.

[251] *See* Tuthill, *supra* note 245.

[252] An early legal case insisting on informed consent frames the problem in these terms: "Every human being of adult years and sound mind has a right to determine what shall be done with his own body; and a surgeon who performs an operation without his patient's consent commits an assault, for which he is liable." Schloendorff v. New York Hosp., 105 N.E. 92, 93 (N.Y. 1914).

[253] *See, e.g.*, U.S. CONST. amend. V ("[N]or shall private property be taken for public use, without just compensation.").

that property in an experiment. But to the extent that a government can enact a policy generally, on the Lockean theory of implicit consent,[254] there should be no ethical bar to the state enacting the policy only against a random set of individuals.

The opposing position flows from the Kantian principle that each person should be treated as an end rather than only as a means.[255] This principle also does not uniquely condemn randomization. Suppose a jurisdiction decides to enact a new universally applicable policy, even though policymakers suspect that it will not be effective but has enough of a chance of success to make it worth trying. If this counts as treating people as means only, then the ethical permissibility of a new policy must be judged excluding from consideration any benefit from the fact that implementation of the policy will produce information about the policy. But many legal regimes, such as patent law and securities law, are justified in part on the basis that they improve information. Information produced by a policy about the policy itself should not be uniquely condemned to irrelevance.

But assuming then that experimentation with universally applicable policies is ethical, is *random* policy experimentation ethical as well? An affirmative case focuses on the benefit of randomization, that it will produce better information than nonrandomized experiments.[256] Although this may at first appear to be a purely consequentialist justification, Robert Veatch argues that subjects of research have a right not to be put "at risk in an unnecessary experiment or one inefficiently designed."[257] The Nuremberg principles on medical experimentation emphasized the importance of experimental design.[258] If universal experimentation is permissible, there is then an *a fortiori* argument that random experimentation must be permissible as well. The difference between the universal and the random experiment is that some people do *not* receive the treatment. Unless there is an equality right to receive the treatment,[259] this difference should not make the experiment more problematic.

---

[254] *See* Peter G. Brown, *Informed Consent in Social Experimentation: Some Cautionary Notes*, *in* ETHICAL AND LEGAL ISSUES OF SOCIAL EXPERIMENTATION, *supra* note 10, at 95-96.

[255] *See* IMMANUEL KANT, GROUNDWORK OF THE METAPHYSIC OF MORALS 429 (H.J. Paton trans., Harper & Row 1964) (1785) ("Act in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means but always at the same time as an end.").

[256] *See supra* Part III.

[257] Robert M. Veatch, *Ethical Principles in Medical Experimentation*, *in* ETHICAL AND LEGAL ISSUES OF SOCIAL EXPERIMENTATION, *supra*, at 21, 37.

[258] *Id.* at 37-38.

[259] *See infra* Part IV.B.1.b.

Medical experimentation itself further supports the argument that if universal policy experiments are permissible, policy experiments must be permissible as well, because medical experiments can generally be viewed equivalently as legal experiments. Subjects in medical experiments who give informed consent presumably would prefer a guarantee of receiving the treatment rather than a chance of a placebo. The status quo is a legal regime that constrains liberty by forbidding distribution of the treatment. Let us assume that the legal prohibition on what Eugene Volokh has called "medical self-defense" is permissible.[260] When the government authorizes a medical experiment,[261] it is effectively authorizing a new legal regime in which patients are permitted to have access to the treatment. The government does not authorize this new legal regime in a universally applicable way, but instead insists on randomization. Only some patients will legally have access to the treatment. It is thus sometimes permissible for new legal policies, including potentially pernicious ones, to be introduced randomly.

This suggests that policy randomization is permissible, at least so long as the group being randomized gives informed consent. The argument for informed consent depends on the legitimacy of legal baselines: both Policy *X* and Policy *Y* are by assumption legally permissible options for policymakers, but if the current policy is *X*, then citizens may only be subject to Policy *Y* if they give informed consent, and vice-versa. The medical experimentation context shows how policymakers can manipulate such baselines. If the baseline were to allow patients to take a medication, then few would consent to being subject to a legal experiment in which they might at random be denied the right to the medicine (equivalently, to a medical experiment in which they might receive a placebo instead).

Those who defend the legitimacy of medical experimentation must either develop an account of which baselines are permissible or allow legal policymakers to play the same game outside the medical context. Existing randomized legal experiments generally allow opt-in to an apparently more favorable treatment. For example, drug offenders may receive the option of participating in an experiment in which they might be randomized to a drug court.[262] This makes

---

[260] *See generally* Abigail Alliance for Better Access to Devel. Drugs v. Eschenbach, 495 F.3d 695 (D.C. Cir. 2007) (en banc) (finding no constitutional right to investigational drugs); Eugene Volokh, *Medical Self-Defense, Prohibited Experimental Therapies, and Payment for Organs*, 120 HARV. L. REV. 1813 (2007) (arguing for a right to medical self-defense).

[261] Government experimentation is necessary for at least some medical experiments. In the United States, the FDA reviews small-scale Phase II trials to determine whether to permit large-scale Phase III trials. *See, e.g.*, Michael D. Greenberg, *AIDS, Experimental Drug Approval, and the FDA New Drug Screening Process*, 3 NYU J. LEGIS. & PUB. POL'Y 295, 305 (1999-2000).

[262] *See, e.g.*, Gottfredson & Exum, *supra* note 111.

experimentation a one-way ratchet, allowing testing within an existing draconian regime of a more lenient alternative, but not allowing testing of more draconian legal approaches. One could test raising the speed limit (by allowing drivers to opt into a program in which they are permitted to drive 10 mph over the limit), but to test lowering the speed limit, policymakers must change the baseline. An insistence on informed consent privileges the status quo legal regime over alternatives, even if neither the status quo nor the alternative applied universally violates any rights.

Might there still be some limits to legal experimentation without informed consent, other than that the policy cannot violate any rights? It is possible to construct extreme hypotheticals. Imagine, for example, a state that wanted more information about the value of public education, and so decided that 1,000 students selected at random would be denied the right to a public education. This hypothetical appears powerful because only a few are randomly selected, because the damage done to the randomly selected is great, because there seems no realistic chance that the policy being tested will be enacted on a universal basis, and because of a view that there may be a right to a public education, even if that right is not constitutionally recognized.[263] The case that the students are being used solely as a means appears particularly strong. But these factors also help explain why there is little chance such a policy would be enacted. Constituents do not want to face some chance of being randomly selected for harsh treatment.

The existence of some problematic experiments that seem highly unlikely to be enacted, but that could be enacted anyway through a change in legal baselines, need not condemn random experimentation on policies that do not violate rights and are genuine candidates for full implementation. Random experimentation might be avoided where the different in treatment, regardless of legal baselines, is sufficiently grave. An experiment on the death penalty, though potentially useful in resolving a difficult and important empirical debate, would presumably be off limits, but that does not necessarily mean that any experiments of alternative sentencing approaches would be banned.[264] At least some experiments, for example those where the unit of

[263] *Compare* Susan H. Bitensky, *Theoretical Foundations for a Right to Education Under the U.S. Constitution: A Beginning to the End of the National Education Crisis*, 86 NW. U. L. REV. 550 (1992) (arguing for such a right), *with* Gregory E. Maggs, *Innovation in Constitutional Law: The Right to Education and the Tricks of the Trade*, 86 NW. U. L. REV. 1038 (1992) (arguing against).

[264] The government, of course, cannot deprive someone of liberty without due process, but a sentencing experiment might well

randomization is at the level of a firm in a regulated industry, should probably be permissible even without informed consent.

b. Equality concerns

Concerns about informed consent focus on the rights of those subject to the experiment. Concerns about equality focus on the rights of those who either are randomly excluded from an experiment or who are assigned to the less desirable of the treatment and control groups. The equality concern is not limited to random experimentation, but extends also to cases in which a government with limited resources distributes those resources at random.[265] For example, governments have used lotteries to distribute scarce low-income housing,[266] rights to immigrate,[267] and positions in magnet and charter schools.[268] Maurice Rosenberg points out that random experimentation may be inevitably in tension with the "equal protection principle . . . that persons subjected to significantly different treatments must be shown to be different in ways that supply a reasonable basis for the differences in treatment."[269] If equal protection were interpreted to prohibit all arbitrary legal differences among similarly situated individuals, then both random experimentation and other programs using random selection to award scarce resources must be eliminated.

There are, however, advantages to using randomization in both these contexts. In the experimental context, randomization has benefits already discussed,[270] and when scarce resources are distributed, randomization ensures that the distribution occurs without favor and in a way that limits rent-seeking for scarce resources.[271] In the United States, the equal protection

---

apply to someone who has already received due process. Indeed, even if randomization is problematic in other contexts, probabilistic criminal punishments may be permissible, if the randomization is conceived and designed as part of the punishment scheme. *See, e.g.*, David Lewis, *The Punishment That Leaves Something to Chance*, 18 PHIL. & PUB. AFF. 53, 58-62 (1999) (defending punishments where the severity is randomized).

[265] Such distribution has generally raised fewer objections than randomization for experimental purposes alone, and as a result experimentation has been particularly feasible in cases in which arbitrary decisions needed to be made in any event. *See* GREENBERG ET AL., *supra* note 3, at 225 (noting that in one experiment, randomization "usually became more acceptable" when officials "recognized that they did not have sufficient funding to serve their entire caseload and, hence, that some mechanism was needed to determine who would be denied services").

[266] *See, e.g.*, Denise Irene Arnold, *Lottery Prize Is Affordable Homes*, N.Y. TIMES, Feb. 7, 1988, at 12 (discussing a local housing lottery).

[267] *See, e.g.*, 8 U.S.C. § 1153(e)(2) (providing for distribution of some visas "strictly in a random order").

[268] *See, e.g.*, Cullen et al., *supra* note 144 (analyzing such a lottery).

[269] Maurice Rosenberg, *The Impact of Procedure-Impact Studies in the Administration of Justice*, LAW & CONTEMP. PROBS., Summer 1988, at 13, 17.

[270] *See supra* Part III.

[271] Rent-seeking can still occur if large numbers of individuals may spend money to enter the lottery. *See, e.g.*, Thomas W.

justification for tolerating both random experimentation and random assignment of government benefits is that there is a rational basis for randomization, and because there is no discrimination against a protected class, no higher standard than rational basis review is necessary. So, in any event, concludes Judge Friendly's opinion in the leading case on this issue.[272] Judge Friendly explained, "The Equal Protection clause does not place a state in a vise where its only choices . . . are to do nothing or plunge into statewide action."[273] A court someday might fail to follow or even overturn this precedent, but it reinforces the plausibility of the legal argument that randomization does not violate the Equal Protection Clause.[274]

But does randomization of legal requirements violate the core principles of equal protection? A full philosophical treatment of this question is beyond this Article's scope, but Ronald Dworkin's treatment of a related issue deserves attention. Dworkin considers the legitimacy of "checkerboard statutes."[275] "Why should Parliament," he asks, "not make abortion criminal for pregnant women who were born in even years but not for those born in odd ones?"[276] Dworkin imagines such a distinction arising from compromise, never considering the possibility that a checkerboard statute might produce useful information. The discussion nevertheless is useful in addressing whether arbitrary distinctions inherently violate equality principles.[277] Dworkin claims that checkerboard statutes offend a principle that he calls "integrity."[278] A jurisdiction enacting such a statute as a compromise "must endorse principles to justify part of what it has done that it must reject to justify the rest."[279] That does not occur with random experimentation, where a single principle, the need to obtain more information, justifies both the treatment and control conditions.[280]

Hazlett & Robert J. Michaels, *The Cost of Rent-Seeking: Evidence from Cellular Telephone License Lotteries*, 59 S. ECON. J. 425 (1993) (analyzing a government lottery that produced 320,000 applications).

[272] Aguayo v. Richardson, 473 F.2d 1090, 1108-10 (2d Cir. 1973).

[273] *Id.* at 1109-10. One commentator has criticized the court for not indicating that its decision would be valid only for as long as the experimental program's value were uncertain. Note, *Reforming the One Step at a Time Justification in Equal Protection Cases*, 90 YALE L.J. 1777, 1783 (1981).

[274] Randomization schemes may sometimes violate other constitutional provisions, however. *See, e.g.*, Delaware v. Prouse, 440 U.S. 648 (1979) (finding random stops of vehicles to check driver's license and registration inconsistent with the Fourth Amendment).

[275] RONALD DWORKIN, LAW'S EMPIRE 178-84 (1986).

[276] *Id.* at 178.

[277] *See id.* at 185 (relating the checkerboard statute issue to conceptions of equality).

[278] *Id.* at 183.

[279] *Id.* at 184.

[280] Another example of Dworkin's reaffirms that arbitrary distinctions are acceptable where not simply the result of legislative

2. Objections to Self-Execution

Checkerboard statutes also suggest an objection to self-execution. If checkerboard statutes represent inappropriate channeling of political compromise, perhaps self-executing statutes should be rejected on the same ground. A justification for self-execution offered above is that it may facilitate compromise.[281] Sometimes, such a compromise will result in an outcome that a policymaker would reject if the policymaker correctly anticipated the result of the experiment. But policymakers routinely make compromises in the face of uncertainty. Legislators unable to reach agreement often leave issues to be resolved by courts[282] or administrative agencies. Self-execution is akin to delegation to experimenters. The outcome might leave some policymakers disappointed, but this does not make the initial decision illegitimate. Lawmakers often enter into incompletely theorized partial agreements when unable to agree on deep justifications for policies.[283]

Once again, we can imagine extreme hypotheticals. Suppose, for example, that conservatives believed (foolishly) that the National League would win the All-Star Game, while liberals fervently believed in the American League. Would it be legitimate for the opposing factions to make important details of legislation contingent on the game's outcome? An experiment could be so irrelevant to the merits of the policy issue that making the experiment self-executing is akin to a coin flip. This would violate the principle that policymakers must engage in reasoned decisionmaking. Although a bright line will be elusive, it seems doubtful that many legislative experiments would be so unrelated to the substantive legal issues as to make this concern salient.

Self-executing experiments will, however, resolve policy debates based on simplified proxies for policy. If a simplified experiment is likely to produce better policy than a more elaborate one, however, that should be sufficient justification. Policymakers have no moral obligation to increase the quantity of societal knowledge at the expense of policy. Admittedly, if

---

compromise: "Suppose we can rescue only some prisoners of tyranny; justice hardly requires rescuing none even when only luck, not any principle, will decide whom we save and whom we leave to torture." *Id.* at 181.

[281] *See supra* Part II.A.2.

[282] As one example, Congress intentionally left unresolved a critical legal issue in passing the Civil Rights Act of 1991. *See* Karen Rosenberg Stein, Comment, *Retroactivity of the Civil Rights Act of 1991: A Decision Not to Decide*, 14 BERKELEY J. EMP. & LAB. L. 275, 276 (1993).

[283] *See generally* Cass R. Sunstein, *Incompletely Theorized Agreements*, 108 HARV. L. REV. 1733 (1995) (introducing the concept of incompletely theorized agreements).

the proxy ex ante seems likely to be so poor that policy will effectively be moving in a random direction, then the case for self-execution is weak. Similarly, if the policy process improves so that it more effectively assimilates expert opinion, more complex experimental designs may be preferable. Even so, self-execution could do little harm, shifting the policy baseline but still permitting policymakers to make changes if subtle experimental results justified them.

V. CONCLUSION

This Article has argued for randomized, self-executing legal experiments. Randomization will generally produce more accurate and easily understood information, and self-execution can facilitate negotiation and ensure that information from experiments affects the law. Such experimentation does not present unique barriers based on concerns of ethics, equality, or accountability. Nonetheless, here are trade-offs. Sometimes, the policy process might incorporate subtle lessons of more complex experiments, so self-execution may not be needed. At other times, effects of nonrandom experimentation might be sufficiently identifiable that randomization is unnecessary. And finally, some policy issues may not lend themselves to experimentation of any form. The absence of randomized experiments with clear effects on policy, however, suggests that experimentation is underused. Conceptualization of experimentation not merely as an academic exercise, but as a standard part of the policy process, is a first step toward identifying and implementing experiments that could potentially improve policy.