University of Pennsylvania Law Review

FOUNDED 1852

Formerly
American Law Register

VOL. 159 | MARCH 2011 | NO. 4

ARTICLE

RANDOMIZING LAW

MICHAEL ABRAMOWICZ,[†] IAN AYRES[††] & YAIR LISTOKIN[†††]

Governments should embrace randomized trials to estimate the efficacy of different laws and regulations. Just as random assignment of treatments is the most powerful method of testing for the causal impact of pharmaceuticals, randomly assigning individuals or firms to different legal rules can help resolve uncertainty about the consequential impacts of law. In this Article, we explain why randomized testing is likely to produce better information than nonrandom evaluation of legal policies. We then offer guidelines for conducting legal experimentation successfully, considering a variety of obstacles, including ethical ones. Randomization will not be useful for all policies, but once government gains better experience with randomization, administrative agencies should pre-

[†] Professor of Law, George Washington University.

[††] William K. Townsend Professor, Yale Law School.

[†††] Associate Professor, Yale Law School.

The authors thank Anthony Vitarelli and John Steele for helpful comments on earlier drafts.

sumptively issue randomization impact statements justifying decisions to implement particular policies. Making the content of law partially contingent on the results of randomized trials will promote ex ante bipartisan agreements, as politicians with different empirical predictions will tend to think that the experiments will support their position.

INTRODUCTION

Legal scholars have debated the impacts of government policy for millenia. In 81 B.C., Chinese scholars argued about the desirability of monopolies in the salt and iron industries in a succession of essays and public debates.[1] These debates were theoretical—with scholars predicting the positive and negative effects of monopolies as compared to a competitive market. Over two thousand years later, theoretical debates over policies remain the norm. But theory alone cannot resolve many policy issues because different theories point in different directions. Scholars attempt to inform these debates by parsing historical data, but regression analysis of policy is fraught with complications. There is little policy variation on many topics of national importance, and the variation that does exist is correlated with many other factors. Empirical policy evaluation often resembles a drug study in which the experimental population does not receive an assigned treatment and instead gets to choose whether to take the medicine or the placebo.

Policymakers and commentators frequently refer loosely to new laws and legal institutions as "experiments,"[2] but in contrast to medical experimentation,[3] these innovations rarely randomly designate treatment and control groups. There have been a handful of exceptions since 1968, including randomized "social experiments" that were

---

[1] *See* A. de G., *The Scholar as Government Consultant: The Great Salt and Iron Debate in Ancient China*, 8 AM. BEHAV. SCIENTIST 4, 4-6 (1965).

[2] *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 Nuremberg and Tokyo war crimes tribunals as "a bold legal experiment"). *See generally* Alan Milner, *Restatement: The Failure of a Legal Experiment*, 20 U. PITT. L. REV. 795 (1959) (characterizing Restatements of the 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").

[3] For a historical discussion of the introduction of randomization into the statistical analysis of medicine, see Tar Timothy Chen, *History of Statistical Thinking in Medicine*, *in* ADVANCED MEDICAL STATISTICS 3, 11-14 (Ying Lu & Ji-Qian Fang eds., 2003). *See also* R.A. Fisher, *The Arrangement of Field Experiments*, 33 J. MINISTRY AGRIC. GR. BRIT. 503, 506-07 (1926) (suggesting the use of random trials in agricultural field experimentation because "[o]ne way of making sure that a valid estimate of error will be obtained is to arrange the plots deliberately at random so that no distinction can creep in between pairs of plots treated alike and pairs treated differently").

performed to assess the impact of government policies.[4] But the legal literature has virtually ignored them. Legal scholars have discussed the results of particular social experiments,[5] and they have commented occasionally that additional social experiments could provide useful information in one field or another.[6] But these legal scholars 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 should take to maximize the chance that experiments will improve policy.

Perhaps as a partial 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.[7] For example, a recent experiment executed under a Medicare statute requiring randomized testing of programs[8] assessed whether telephone

---

[4] A doctoral student, Heather Ross, developed the idea for an experiment on the effect of a negative income tax and then received governmental funding for her experiment. The experimental results are reported in three volumes. *See generally* DAVID KERSHAW & JERILYN FAIR, 1 THE NEW JERSEY INCOME-MAINTENANCE EXPERIMENT: OPERATIONS, SURVEYS, AND ADMINISTRATION (1976); 2 THE NEW JERSEY INCOME-MAINTENANCE EXPERIMENT: LABOR-SUPPLY RESPONSES (Harold W. Watts & Albert Rees eds., 1977); 3 THE NEW JERSEY INCOME-MAINTENANCE EXPERIMENT: EXPENDITURES, HEALTH, AND SOCIAL BEHAVIOR; AND THE QUALITY OF THE EVIDENCE (Harold W. Watts & Albert Rees eds., 1977). For useful summaries of the experiment, see DAVID GREENBERG ET AL., SOCIAL EXPERIMENTATION AND PUBLIC POLICYMAKING 111-64 (2003); Frank P. Stafford, *Income-Maintenance Policy and Work Effort: Learning from Experiments and Labor-Market Studies*, *in* SOCIAL EXPERIMENTATION 95, 111 tbl.3.5 (Jerry A. Hausman & David A. Wise eds., 1985).

[5] *See, e.g.*, Machaela M. Hoctor, Comment, *Domestic Violence as a Crime Against the State: The Need for Mandatory Arrest in California*, 85 CALIF. L. REV. 643, 655-57 (1997) (commenting on a Minneapolis experiment, in which police, assuming probable cause, assigned at random the arrest of alleged domestic violence perpetrators, and noting that suspects who were arrested had the lowest rate of recidivism).

[6] *See, e.g.*, Bernard E. Harcourt, *Post-Modern Meditations on Punishment: On the Limits of Reason and the Virtues of Randomization* (A Polemic and Manifesto for the Twenty-First Century), 74 SOC. RES. 307, 328-30 (2007) (proposing randomization in several areas, such as criminal justice, where it could be used to set the length of prison sentences); Laurens Walker, *Perfecting Federal Civil Rules: A Proposal for Restricted Field Experiments*, LAW & CONTEMP. PROBS., Summer 1988, at 67, 72-77 (proposing randomized experiments on rules of civil procedure).

[7] *See* GREENBERG ET AL., *supra* note 4, at 26 ("[M]ost social experiment test programs are targeted at persons or families who are somehow disadvantaged, particularly in terms of having low incomes.").

[8] Medicare Prescription Drug, Improvement, and Modernization Act of 2003 § 721(b)(1), 42 U.S.C. § 1395b-8(b)(1) (2006) (requiring "development, testing, and evaluation of chronic care improvement programs using randomized controlled trials").

contact by nurses to at-risk Medicare patients would reduce program costs.[9] Another class of randomized studies evaluated criminal justice policies.[10] A rare exception to these two areas has been a set of experiments on electricity pricing.[11] Experiments have almost never varied the legal rights and obligations of ordinary citizens or entities in areas such as securities law or taxation.[12] Instead, experiments have focused on the possible provision of new services or on those who might be thought of as forfeiting rights by committing crimes.

This Article advances the case for randomizing law, including the legal rights and obligations expressed in statutes and regulations.[13] Randomized experiments have the potential not only to be governmentally funded academic exercises, but also to serve as integral components of the legal process. In this Article, we argue that government should embrace randomized trials of statutes and regulations as a tool for testing the effectiveness of those laws. Just as random assignment of treatments is the most powerful method of testing for the causal impact of pharmaceuticals, random assignment of individuals, firms, or jurisdictions to different legal rules can help resolve uncertainty about the consequences of laws and regulations.

Beyond endorsing randomized legal experimentation in areas where such experiments have not generally been contemplated, this Article considers how the policy process should change to accommodate randomized experimentation. Administrative law, we argue,

---

[9] *See* NANCY MCCALL ET AL., RTI INT'L, EVALUATION OF PHASE 1 OF MEDICARE HEALTH SUPPORT (FORMERLY VOLUNTARY CHRONIC CARE IMPROVEMENT) PILOT PROGRAM UNDER TRADITIONAL FEE-FOR-SERVICE MEDICARE 1-5 (2007) (reporting preliminary results of the programs); Reed Abelson, *Medicare Finds How Hard It Is to Save Money*, N.Y. TIMES, Apr. 7, 2008, at A1 (describing one such program).

[10] *See generally* David P. Farrington & Brandon C. Welsh, *A Half Century of Randomized Experiments on Crime and Justice*, 34 CRIME & JUST. 55 (2006) (providing an overview of randomized criminal justice experiments from 1957 to 2004).

[11] *See* Dennis J. Aigner, *The Residential Electricity Time-of-Use Pricing Experiments: What Have We Learned?* ("The purpose of the present paper is to consider the empirical results available so far from the DOE experiments in light of design and analysis concerns . . . ."), *in* SOCIAL EXPERIMENTATION, *supra* note 4, at 11, 12; *see also* 1 RESEARCH TRIANGLE INST., ANALYTICAL MASTER PLAN FOR THE ANALYSIS OF THE DATA FROM THE ELECTRIC UTILITY RATE DEMONSTRATION PROJECTS 1 (1978) (analyzing projects designed to "evaluate experimentally the effects of time-of-use pricing of electricity for residential customers").

[12] For an exception that we propose to extend, see *infra* subsection IV.C.1.a.

[13] The possibility that judge-made legal rules could be subjected to randomized testing is beyond the scope of this Article. Such testing could be implemented by legislatures to the extent that statutes can preempt common law rulemaking. But, more speculatively, one might imagine courts themselves conducting prospective randomized control experiments to gather evidence on the most appropriate resolution in a case.

should accept decisions by agencies to randomize policies and perhaps even be more deferential to policy decisions made after a process of experimentation. Ultimately, the executive branch could make formalized consideration of randomized control trials as central to the regulatory process as formalized consideration of the costs and benefits of regulations. If experimentation begins to occur with sufficient frequency in agencies, Congress or other legislatures might themselves initiate experiments more frequently. The possibility of experimentation may reduce legislative disagreement. Where disagreements are truly empirical, partisans on both sides of an issue may believe that they would benefit from experimentation. A self-executing experiment can, in effect, serve to resolve a bet among competing legislative factions, with the experimental outcome automatically affecting the content of the legislation. Meanwhile, if a legal culture of randomization developed sufficiently, a legislator's refusal to endorse an experiment might be interpreted as evidence that the legislator's empirical claims about a policy mask some other agenda.

The Article proceeds as follows. Part I lays out the affirmative case for randomized control trials and describes our central proposal. Part II describes the problems of nonrandom evaluation of legal policies. Conventional regression analysis is subject to problems, including omitted variable bias, publication bias, and misspecification. Part III discusses potential problems and pitfalls of randomized policy experiments, as well as responses to these complications. Some of these problems involve challenges of interpreting even randomized legal experiments, though, in general, randomization should make interpretation somewhat easier. The more challenging problems from the perspective of policy implementation are that randomized legal policy may be costly or raise ethical concerns. Finally, Part IV offers some guidelines for legal experimentation, including specific recommendations for legislatures and administrative agencies, and it then describes specific applications in which randomization seems especially likely to be fruitful.

I. THE POWER OF RANDOMIZED CONTROLS

The idea that randomization could be used to create a quality control group has existed since 1925, when Ronald Fisher, the father of modern statistics, suggested using random assignments in research involving agricultural trials that arose out of his work at the Ro-

thamsted Experimental Station.[14] In his 1935 book, *The Design of Experiments*, Fisher explained the power of the technique with the arresting example of a "[l]ady [who] declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup."[15] Fisher proposed mixing eight cups of tea—four with milk first and four with milk last—and "presenting them to the subject for judgment in a random order."[16]

Intentionally interjecting uncertainty into the experimental design could have the perverse effect of enhancing the ability of a researcher to control the experiment. As David Harrington has noted,

> [i]n one of the delightful ironies of modern science, the randomized trial "adjusts" for both observed and unobserved heterogeneity in a controlled experiment by introducing chance variation into the study design. If interventions for patients are chosen by chance, then the law of large numbers implies that the average values of patient characteristics should be roughly equal in the intervention groups.[17]

Randomization itself produces the controlled environment in which a similar group may provide a source of comparison. Of course, randomization does not mean that the control and treatment groups will be identical. If we looked at, for example, the heights of people in each group, we would see the normally distributed bell curve. But the point is that we would see similarly shaped bell curves of heights in both groups. The law of large numbers assures that as the size of the group increases, the mean of both groups will both converge on the population mean. But random assignment means that beyond the mean, the *distribution* of both groups with regard to every characteristic (save the treatment itself) will become increasingly identical as the sample size increases. Instead of trying to establish identical control pairs—which on a pair-wise basis are identical on every nontreatment dimension—random assignment creates groups that have statistically similar distributions on every nontreatment dimension. Since the distribution of height (or any other characteristic) is the same in both

---

[14] *See* R.A. FISHER, 3 STATISTICAL METHODS FOR RESEARCH WORKERS 203-09 (1925) (describing this experiment); *see also* IAN AYRES, SUPER CRUNCHERS 46-80 (2007) (discussing the power of randomization as a tool for business and the government).

[15] RONALD A. FISHER, THE DESIGN OF EXPERIMENTS 11 (Hafner Publ'g Co., 6th ed. 1951) (1935).

[16] *Id.*

[17] David P. Harrington, *The Randomized Clinical Trial*, 95 J. AM. STAT. ASS'N 312, 312 (2000).

the control and the treatment groups, we can attribute any differences in the *average* group response to the difference in treatment.[18]

Fisher's breakthrough was realizing that randomization could do a better job of producing a controlled experiment than would non-randomized controls. Fisher went so far as to argue that randomization produced better controls than could *ever* be achieved by physically matching the nontested attributes. In discussing his "Lady and the Tea" problem, Fisher explained:

> It is no sufficient remedy to insist that "all the cups must be exactly alike" in every respect except that to be tested. For this is a totally impossible requirement in our example, and *equally in all other forms of experimentation.* In practice it is possible that the cups will differ perceptibly in the thickness or smoothness of their material, that the quantities of milk added to the different cups will not be exactly equal, that the strength of the infusion of tea may change between pouring the first and the last cup, and that the temperature also at which the tea is tasted will change during the course of the experiment.[19]

For Fisher, some attributes of an experiment were beyond a researcher's ability to physically control through experimental design. Some causal attributes, for example, may not be observable. But randomization as a control assures that sufficiently large control and treatment groups will be similar even with regard to attributes that are unobservable to the researcher.

The earliest formal randomized drug trial on humans took place in the late 1940s, when Austin Bradford Hill ran the first clinical trial testing the effectiveness of the antibiotic streptomycin in treating tuberculosis.[20] By 1962, the use of random controlled trials had become so prevalent that Congress amended the Food, Drug, and Cosmetic Act to mandate the use of "adequate and well-controlled investigations, including clinical investigations, by experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved."[21]

---

[18] For more information on the logic of clinical trials, see generally *id.*

[19] FISHER, *supra* note 15, at 18 (emphasis added).

[20] For more information about the trial and its methodology, see Streptomycin in Tuberculosis Trials Comm., Med. Research Council, *Streptomycin Treatment of Pulmonary Tuberculosis*, 2 BMJ 769 (1948).

[21] Pub. L. No. 87-781, § 102(c), 76 Stat. 780, 781 (1962) (codified as amended at 21 U.S.C. § 355(d) (2006)). *See generally* Karen Baswell, Note, *Time for a Change: Why the FDA Should Require Greater Disclosure of Differences of Opinion on the Safety and Efficacy of Approved Drugs*, 35 HOFSTRA L. REV. 1799, 1832 (2007) (arguing for "increased transparency and disclosure from the pharmaceutical and pharmaceutical research industries").

Since 1970, randomized clinical trials have been a critical part of the U.S. Food and Drug Administration's analysis of applications.[22]

Given the considerable benefits of randomized policy experiments, we propose that the government systematize and expand experimentation. Before enacting legislation, legislators should consider conducting an experiment of the new policy. Administrators should also adopt widespread experimentation. Just as cost-benefit analyses and "environmental impact statements" (EIS) are necessary steps in the formation of numerous regulations and policies,[23] so too should "randomization impact statements" (RIS) become part of the policy planning landscape. Randomized studies should not be an absolute prerequisite for legal change. A norm to randomize or explain why randomization could not be undertaken, however, would help discipline regulators to take evidence-based lawmaking more seriously. Whenever a new regulation is put forward, the relevant agency should be presumptively required to present an RIS with the contents described in this Part. We discuss later the details of implementation (including when an

---

[22] *See* Abigail Alliance for Better Access to Developmental Drugs v. Von Eschenbach, 495 F.3d 695, 697, 706 n.12 (D.C. Cir. 2007) (en banc) (considering whether "terminally ill patients [have] a right of access to experimental drugs that have passed limited safety trials but have not been proven safe and effective" and noting that "'[t]he history of the effectiveness requirement in drug regulation is inextricably linked to the advent of the randomized, controlled clinical trial as the cornerstone of medical research'" (quoting Jennifer Kulynych, *Will FDA Relinquish the "Gold Standard" for New Drug Approval? Redefining "Substantial Evidence" in the FDA Modernization Act of 1997*, 54 FOOD & DRUG L.J. 127, 131 (1999))); 21 C.F.R. § 314.50(d)(5) (2010) (describing the required presentation of controlled and uncontrolled clinical trials in an application for FDA approval); *see also* Charles J. Walsh & Alissa Pyrich, *Rationalizing the Regulation of Prescription Drugs and Medical Devices: Perspectives on Private Certification and Tort Reform*, 48 RUTGERS L. REV. 883, 889 (1996) (suggesting a plan to delegate some of the FDA's regulatory authority to private bodies in order to increase efficiency); *cf.* 40 C.F.R. § 799.9420(d)(1)(iv)(D) (2009) (mandating randomized testing of toxic substances in rodents).

[23] The National Environmental Policy Act of 1969, 42 U.S.C. § 4332(2)(C) (2006), requires an EIS for any "major Federal action[] significantly affecting the quality of the human environment." The purpose of the EIS is to improve agency decisionmaking by requiring "detailed information concerning significant environmental impacts." Robertson v. Methow Valley Citizens Council, 490 U.S. 332, 349 (1989). Executive Order Number 12,866 states that "[i]n deciding whether and how to regulate, agencies should assess all costs and benefits of available regulatory alternatives, including the alternative of not regulating." Exec. Order No. 12,866, 3 C.F.R. 638, 639 (1994), *reprinted as amended in* 5 U.S.C. § 601 (2006). "The objectives of this Executive Order are to enhance planning and coordination . . . to reaffirm the primacy of Federal agencies in the regulatory decision-making process; to restore the integrity and legitimacy of regulatory review and oversight; and to make the process more accessible and open to the public." *Id.* at 638.

agency may proceed to regulate without an RIS).[24] The new norm, however, should be the presentation of data from a randomized policy experiment.

II. THE PROBLEMS OF NONRANDOM EVALUATION

This Part explores the advantages of randomized studies by reviewing recurring weaknesses in alternative modes of evaluation. This analysis responds to an anticipated counterargument—that randomized studies are unnecessary—because statistical and econometric techniques can be used to estimate policy effects reliably. Even when the most advanced techniques are employed, however, nonrandom analyses will generally leave more uncertainty than random analyses. Any statutory change is experimental in that it creates a new legal regime, allowing comparison to the world under the prior regime. Indeed, it is common for proponents and neutral commentators to describe such a change as "an experiment."[25] 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 it is possible to link cause and effect. 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.[26] Those who do not have the time, inclination, or ability to probe the evidence cannot then easily discern the truth.[27]

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 correlations between new policies and other variables need not imply causation. It will thus almost always be relatively easy for partisans to find some basis on which to develop misleading theories or else to offer critiques of relatively robust results. This Part explains why, even using data from numerous jurisdictions, the technique of convent-

---

[24] *See infra* Section IV.B.

[25] *See supra* note 2 and accompanying text.

[26] *See* Bryan Keefer, *Tsunami*, COLUM. JOURNALISM REV., July–Aug. 2004, at 18, 18-23 (discussing reporters' reluctance to take sides on issues of public controversy during elections).

[27] *Cf.* Scott Brewer, *Scientific Expert Testimony and Intellectual Due Process*, 107 YALE L.J. 1535, 1552-53 (1998) (describing a similar problem when judges and jurors try to assess scientific evidence beyond their competence).

ional multiple regression analysis, in which the policy of interest forms an independent variable, may produce inaccurate results. This Part, of course, is not intended to provide a comprehensive examination of the uses and limits of statistical analysis.[28] Section II.B comments on the difficulties of improving the law by using the states as policy laboratories without randomization.

A. *Conventional Regression Analysis*

1. 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.[29] This result could mean that sex education encourages students to have more sex, but it also could reflect the fact that school districts with high rates of student sexual activity respond to these rates by offering sex education. Statisticians overcome this problem by adding control variables for the characteristics of the students, such as family income, parents' education, and religion, as well as for the characteristics of the community, such as whether it is rural or urban and the region of the country in which it is located.[30] If those variables exhaust all nonrandom factors contributing to community and family decisions about sex education, then the coefficient on the sex education variable would successfully represent the effect of random variation on whether students are exposed to sex education. But if there is an omitted variable that correlates with both the community decision to offer sex education and the individual decision to have sex, the coefficient will be biased.

Even careful researchers cannot easily avoid this problem (and it can be exploited by researchers who hope to establish a particular result). There are at least two reasons it is difficult to avoid bias. First, the available data may be incomplete. Even if there are strong theo-

---

[28] For a useful overview of regression analysis, see generally WILLIAM MENDENHALL & TERRY 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 & Gary King, *The Rules of Inference*, 69 U. CHI. L. REV. 1, 14-19 (2002).

[29] For a study on the relationship between sexual activity and sex education, see Deborah Anne Dawson, *The Effects of Sex Education on Adolescent Behavior*, 18 FAMILY PLANNING PERSP. 162, 162 (1986).

[30] *See, e.g.*, *id.* at 170 tbl.9 (listing control variables and their respective effects on the likelihood of premarital pregnancy).

retical reasons to believe that parental education is an important variable, it may be impossible to develop a measure that fully accounts for the parents' educational level.[31] For example, a measure indicating whether someone's mother graduated from high school would seem to imply that all high-school dropouts are alike and that 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 which variables 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 particular governmental programs. For example, Julie Berry Cullen and others analyzed the effect of school-choice lotteries, whose winners would be allowed to attend particular schools.[32] They noted that, according to some studies, students who won the school-choice lotteries appeared to do better than students who did not enter the lotteries.[33] Competing explanations for this result included that lottery winners were allowed to attend better schools and that more motivated students are likely to self-select into the lottery.[34] In the absence of variables fully capturing student motivation, a regression would tend to inflate the apparent effects of the schools on student performance. Indeed, the Cullen study showed that students who won the school-choice lotteries performed no better than students who entered but lost the same lotteries.[35] So it was student motivation, and not school quality, that caused the difference in performance between school-lottery winners and nonlottery entrants. Though not created for the purpose of facilitating data analysis, the lottery pro-

---

[31] Dawson's study used a binary variable indicating whether the mother had at least twelve years of education. *Id.*

[32] Julie Berry Cullen et al., *The Effect of School Choice on Student Outcomes: Evidence from Randomized Lotteries* (Nat'l Bureau of Econ. Research, Working Paper No. 10113, 2003).

[33] *See id.* at 1 (discussing such findings in previous observational studies).

[34] *See id.* (noting that the studies producing such results "suffer potentially from important selection bias since the students who take advantage of school choice are unlikely to be representative of students more generally"); *see also, e.g.*, Caroline M. Hoxby & Sonali Murarka, Methods of Assessing the Achievement of Students in Charter Schools 24 (Sept. 28, 2006) (unpublished manuscript), *available at* http://www.vanderbilt.edu/schoolchoice/conference/papers/Hoxby-Murarka_2006-DRAFT.pdf (noting possible variables that increase self-selection bias in students who apply to charter schools, such as "prior achievement" or "parental motivation").

[35] Cullen et al., *supra* note 32, at 23.

duced random assignments that allowed the researchers to avoid omitted variable bias.[36]

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 reality can be shown by comparing the results of randomized experiments with the results of nonrandomized statistical analysis. Paul Glewwe and others conducted separate prospective randomized and retrospective nonrandomized studies of whether making "flip charts"—large visual aids illustrating concepts in science, health, and mathematics—available to students in Kenya increased test scores.[37] The retrospective studies showed that flip charts increased test scores, while the randomized studies revealed no effect.[38] 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.[39] The forces that lead jurisdictions or 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.

Some statistical techniques, such as instrumental-variable and regression-discontinuity studies, seek to take advantage of naturally occurring randomness.[40] A full discussion of these techniques is beyond

---

[36] Hoxby & Murarka, *supra* note 34, at 24 ("[R]andomization over a large number of students who apply to a charter school eliminates all forms of self-selection bias . . . .").

[37] Paul Glewwe et al., *Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya* (Nat'l Bureau of Econ. Research, Working Paper No. 8018, 2000).

[38] *Id.* at 18.

[39] *See id.* at 8 (concluding that "adding flip charts raises test scores by about 20 percent of a standard deviation in flip-chart subjects").

[40] Consider, for example, Steven Levitt's attempt to determine the effect of police presence on crime. *See* Steven D. Levitt, *Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime*, 87 AM. ECON. REV. 270 (1997). 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 "endogen[ous]" and may depend in part on expectations of whether crime is likely to increase. *Id.* at 270. Levitt developed a regression predicting the change in the number of police officers based on factors including whether a mayoral election was scheduled, *id.* at 277, 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. *See id.* at 271 (noting that the "instrument employed" was "the timing of mayoral and gubernatorial elections"). For another example of an instrumental-variables study, see Jonathan Klick & Alexander Tabarrok, *Using Terror Alert Levels to Estimate the Effect of Police on Crime*, 48 J.L. & ECON. 267 (2005), which uses terror-alert levels in Washington,

the scope of this Article, but these approaches are often inferior to randomized control trials. With instrumental-variables studies, there may be some subjectivity in the choice of instruments. Although there are statistical tests that can be used to assess the validity of instruments,[41] one can still argue about whether specific instruments are the best available. Meanwhile, with regression-discontinuity studies, there may be some subjectivity in assessing whether groups on either side of a discontinuity are truly comparable.[42] Casual empiricism, in any event, suggests that such studies require sufficient analytical judgment such 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., "when cell phone providers systematically transition from 'peak' to 'off-peak' pricing."[43] Call volumes increase discontinuously around the 9 p.m. threshold, but there has been no recent increase in car accidents immediately after 9 p.m. compared to the period before cell companies began to offer free calling after 9 p.m.[44] Nonetheless, policymakers have continued to proclaim driving while talking on a cell phone to be dangerous.[45] Instrumental-variables and

---

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. For an overview of regression-discontinuity studies, see Jinyong Hahn et al., *Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design*, 69 ECONOMETRICA 201 (2001). An example of such a study is M. Keith Chen & Jesse M. Shapiro, *Do Harsher Prison Conditions Reduce Recidivism? A Discontinuity-Based Approach*, 9 AM. L. & ECON. REV. 1 (2007), which uses the variability in prison conditions as a function of offense levels to analyze recidivism rates.

[41] *See* J.A. Hausman, *Specification Tests in Econometrics*, 46 ECONOMETRICA 1251, 1251 (1978) (proposing a "general form of specification test").

[42] For example, Chen and Shapiro compared the demographics of two similar populations and found some significant demographic differences. *See* Chen & Shapiro, *supra* note 40, at 16. 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," *id.* at 17, including binary controls for the cutoffs, *id.* at 19-20. 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.

[43] Saurabh Bhargava & Vikram Pathania, *Driving Under the (Cellular) Influence: The Link Between Cell Phone Use and Vehicle Crashes* 3 (AEI-Brookings Joint Ctr. for Regulatory Studies, Working Paper No. 07-15, 2007), *available at* http://ssrn.com/abstract=1089081.

[44] *Id.* at 4-5.

[45] *Cf.* Mike Stuckey, *Hands-Free Phones Are Lifesavers, Study Says*, MSNBC.COM, May 13, 2008, http://www.msnbc.msn.com/id/24580099 (discussing a study "predicting that banning the use of hand-held phones by U.S. drivers could save thousands of lives each year").

regression-discontinuity studies do not necessarily have a greater impact on the policy process than other studies, even with respect to the issues for which they are feasible.

2. Publication Bias and Misspecification

Statistically significant results can also be spurious as a result of publication bias. Finding a statistically significant outcome at the generally accepted 0.05 level[46] means that there is a five-percent chance that an outcome at least as extreme would have occurred by pure chance if the null hypothesis were true.[47] If, for example, researchers test 100 propositions that in fact are all false, about five of these tests may produce statistically significant results, and these mistaken results will be the most publishable.[48] Meanwhile, insignificant findings provide little support for the truth of the corresponding null hypotheses. Insignificant findings are difficult to publish, but they may be publishable when they are counterintuitive. There is a relatively high probability, however, that rare counterintuitive failures to reject the null hypothesis are the result of chance.[49]

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,[50] which variables to include, what transformations to apply to the variables,[51] and which observations to include.[52] Especially within social science, researchers do not neces-

---

[46] *See* Fisher, *supra* note 3, at 504 ("A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give [a five-percent] level of significance." (emphasis omitted)).

[47] An example of a null hypothesis would be that in a regression equation's estimate of the true relationship, the coefficient for one of the independent variables is zero, which indicates that, after controlling for other variables, there is no relationship between the dependent variable and that independent variable.

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

[49] *See* J. Bradford De Long & Kevin Lang, *Are All Economic Hypotheses False?*, 100 J. POL. ECON. 1257, 1258 (1992) (discussing "an approach that allows [its authors] to measure the fraction of *unrejected* null hypotheses that are, in fact, *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.

[50] *See, e.g.*, WILLIAM H. GREENE, ECONOMETRIC ANALYSIS 116-144 (5th ed. 2003) (discussing the "functional form of the regression model").

[51] *See id.* at 347 n.11 (considering the possibility of nonlinear specifications).

[52] 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 a researcher identifies an observation as an outlier, she will run a regression both

sarily settle on regression specifications in advance, but instead "pretest" data to determine which results to report.[53] Considering a large number of regression specifications may help researchers develop nuanced conclusions, but researchers will generally be more likely to report results producing statistical significance.[54] Laboratory experiments are also subject to publication bias, but other researchers can attempt replication. Social science researchers cannot rerun history.[55]

Social scientists can, however, seek to assess the robustness of published results, but often there will be some subjectivity involved in determining whether a study's results are sufficiently robust to justify causal inferences. A recent example was John Donohue and Justin Wolfers' scrutiny of studies purporting to show the death penalty's deterrent effects.[56] For example, they criticized a study by Hashem Dezhbakhsh and Joanna Shepherd,[57] focusing first on a cross-sectional analysis of homicide trends in states that either abolished or adopted the death penalty, as well as states that never had a death penalty.[58] Donohue and Wolfers argue that the same general trends existed in states

---

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, ROBUST REGRESSION AND OUTLIER DETECTION 10 (2003) (describing regression techniques that account for the large influence of outliers).

[53] *See* 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.").

[54] The traditional *t*-statistic will be inaccurate when researchers test numerous regression specifications and then focus on only those whose *t*-statistics appear to indicate statistically significant results. *See, e.g.*, Dmitry Danilov & Jan R. Magnus, *Forecast Accuracy After Pretesting with an Application to the Stock Market*, 23 J. FORECASTING 251, 258 (2004) (deciding which model to use depending on whether the *t*-statistic is significant). For a discussion of the origin and formation of the *t*-statistic, see ARTHUR M. GLENBERG & MATTHEW E. ANDRZEJEWSKI, LEARNING FROM DATA 243 (3d ed. 2008).

[55] *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").

[56] *See* John J. Donohue & Justin Wolfers, *Uses and Abuses of Empirical Evidence in the Death Penalty Debate*, 58 STAN. L. REV. 791, 840 & fig.10 (2005) (finding a "statistically significant correlation" between the estimated coefficients and standard errors reported within most studies).

[57] *See* Hashem Dezhbakhsh & Joanna M. Shepherd, *The Deterrent Effect of Capital Punishment: Evidence from a "Judicial Experiment"* 13-15 (Emory L. & Econ. Paper Series, Working Paper No. 04-04 & Emory Univ. Econ. Working Paper No. 03-14, 2004), *available at* http://ssrn.com/abstract=432621 (analyzing the "effect of both suspending and reinstating the death penalty on state murder rates" and concluding that "states reinstating their death penalty experience a drop in murder rates").

[58] Donohue & Wolfers, *supra* note 56, at 800-04.

that had not changed their death penalty policies, and Donohue and Wolfers reanalyzed the data with a difference-in-differences approach.[59] This revised analysis produced statistically insignificant results.[60]

Such a conclusion does not mean that every empirical question will yield an uncertain answer. But the death penalty is hardly the only debate about which scholars 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[61] and whether statutes allowing citizens to carry concealed handguns lower violent crime rates.[62] Whatever the merits of the various arguments, academics and policymakers have not reached a consensus on these questions. Even if the median voter or decisionmaker would be swayed by empirical results, it will not be easy to determine what results to believe.

Just as in nonrandom evaluations, publication bias is also a danger in randomized studies.[63] But there is less room for identifying alternative empirical specifications given the centrality of the random variable in randomized trials. As Esther Duflo has noted, in retrospective studies, "the researchers or evaluators define their own comparison group [ex post], and thus may be able to pick a variety of plausible comparison groups,"[64] but in a randomized study, the treatment and comparison groups will generally be clearly defined at the outset of the study. There is still some danger that researchers will decide not to

---

[59] *See id.* at 801-02 (arguing that "focusing *only*" on states that altered their policies risks "confounding the effects of changes in capital punishment laws with broader forces").

[60] *See id.* at 802 (finding that a difference-in-differences approach produces "no evidence that the death penalty affects homicide rates").

[61] *See* John J. Donohue III & Steven D. Levitt, *The Impact of Legalized Abortion on Crime*, 116 Q.J. ECON. 379, 414 (2001) (presenting evidence suggesting that "legalized abortion is a primary explanation for the large drops in murder, property, crime, and violent crime").

[62] *See* JOHN R. LOTT, JR., MORE GUNS, LESS CRIME 159 (1998) ("Allowing citizens without criminal records or histories of significant mental illness to carry concealed handguns deters violent crimes . . . .").

[63] 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.*, Erick H. Turner et al., *Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy*, 358 NEW ENG. J. MED. 252, 253-56 (2008) (analyzing which reviews of antidepressant agents submitted to the FDA were subsequently published and observing a "bias toward the publication of positive results").

[64] Esther Duflo, *Scaling Up and Evaluation*, *in* ANNUAL WORLD BANK CONFERENCE ON DEVELOPMENT ECONOMICS 2004: ACCELERATING DEVELOPMENT 341, 353 (François Bourguignon & Boris Pleskovic eds., 2004).

publish, but that danger is considerably reduced when governmental institutions have sponsored the research by supporting the randomization of policy and when a particular set of researchers has promised to analyze the effects of the experiment. Indeed, governments can virtually eliminate the risk of nonpublication by requiring publication of experimental results as a condition of funding.[65]

B. *The Laboratory of the States Reconsidered*

For statistical research to influence policy rather than merely serve as a sound bite, it must 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 a kind of laboratory to test policies.[66] Susan Rose-Ackerman, for example, has shown that federalism may insufficiently promote experimentation for numerous reasons,[67] including because one state may hope to free ride on the activities of other governments.[68] Edward Rubin and Malcolm Feeley have similarly noted that experimentation sometimes may be expensive and ultimately not beneficial

---

[65] For an argument in favor of clinical trial registries that require drug companies to release both positive and negative studies, see James M. Wood & Roxanne M. Gariby, *Hoarding Away Science: Towards a More Transparent View of Health and Online Registries for Independent Postmarket Drug Research*, 60 FOOD & DRUG L.J. 547 (2005).

[66] Justice Brandeis articulated the classic statement of this theory in his dissent in *New State Ice Co. v. Liebmann*, 285 U.S. 262 (1932). He wrote, "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." *Id.* at 262 (Brandeis, J., dissenting). For a discussion of this justification for federalism, see Ann Althouse, *Vanguard States, Laggard States: Federalism and Constitutional Rights*, 152 U. PA. L. REV. 1745, 1750-76 (2004).

[67] *See* Susan Rose-Ackerman, *Risk Taking and Reelection: Does Federalism Promote Innovation?*, 9 J. LEGAL STUD. 593, 615 (1980) (stating that federalism produces only "weak effects" in promoting innovation).

[68] *See id.* at 605 ("The better other governments are expected to do, the less incentive any politician has to initiate projects."). The possibility that there might be insufficient incentives to innovate is apparent even in areas in which state competition has generally been trumpeted, such as corporate governance law. *See* Michael Abramowicz, *Speeding up the Crawl to the Top*, 20 YALE J. ON REG. 139, 143 (2003) (suggesting that "the existence of state competition is not enough to guarantee an optimal level of innovation"). *But see* Roberta Romano, *The States as a Laboratory: Legal Innovation and State Competition for Corporate Charters*, 23 YALE J. ON REG. 209, 246 (2006) (arguing that states are, in fact, "effective laborator[ies]" in the "corporate chartering context").

for the experimenter,[69] suggesting that centralized coordination may be needed to take full advantage of federalism.[70]

Yet, at least sometimes, states do change their policies—and take risks in doing so—in the hope of achieving informational benefits.[71] As Barry Friedman notes, states may innovate for a variety of reasons, quite apart from any desire to engage in "experimentation."[72] These state innovations serve at least a crude experimentation function.[73] Commentators may observe that one state's approach to a particular issue, such as health care reform, has gone particularly poorly or well, and this observation may influence decisionmaking in other states.[74] Federalism, however, does not easily facilitate a scientific approach to experimentation. The difficulties that social scientists and especially policymakers face in assessing the results of state innovations contribute to the inaptness of the states-as-laboratories metaphor.

Still, federalism may be more conducive to experimentation than the alternatives. Previous commentators have noted that randomized experiments are much more common in North America than in the rest of the world,[75] and they have speculated that federalism may help explain this.[76] In any event, the mere existence of different jurisdic-

---

[69] *See* Edward L. Rubin & Malcolm Feeley, *Federalism: Some Notes on a National Neurosis*, 41 UCLA L. REV. 903, 925 (1994) ("[I]ndividual states will have no incentive to invest in experiments that involve any substance or political risk . . . .").

[70] *See id.* at 926 (noting that absent "coordinat[ion] by a centralized authority . . . state-initiated experiments are unlikely to be truly useful to other states because of more specific, technical variations" among the states).

[71] *Cf.* FERC v. Mississippi, 456 U.S. 742, 788 (1982) (O'Connor, J., concurring in the judgment in part and dissenting in part) ("[S]tate innovation is no judicial myth.").

[72] Barry Friedman, *Valuing Federalism*, 82 MINN. L. REV. 317, 398-99 (1997) ("'Innovation' might have been a better word choice for Justice Brandeis than 'experimentation,' saving us all a lot of bother.").

[73] 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 2, at 345 (explaining that 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").

[74] *See, e.g.*, Sara Rosenbaum, *Mothers and Children Last: The Oregon Medicaid Experiment*, 18 AM. J.L. & MED. 97, 126 (1992) (concluding that Oregon's series of Medicaid initiatives is an experiment that "this nation cannot afford to conduct" because it "fails the basic test of civility").

[75] *See* GREENBERG ET AL., *supra* note 4, at 38 (noting as an exception that the Netherlands tested an unemployment-counseling program).

[76] *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.*

tions 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. 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.[77]

III. CAVEATS: LIMITS OF RANDOMIZATION STUDIES

A. *Interpretive Problems*

Advocates of randomized studies have emphasized that only these types of studies can establish causality with high confidence. For example, Esther Duflo has argued:

> [W]hile 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 . . . .[78]

But what "is relatively easy to ascertain" may 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 law enacting the legal experiment. This Section suggests that this requirement may be explained because legal experiments will not generally be double-blind, because it may be difficult to generalize from the experimental context to the ultimate policy context, and because randomization may be imperfect.

1. Non-Double-Blind Randomization

The purest form of a randomized experiment includes informational control over both the researcher and the subjects. In double-blind experiments, for example, neither the researchers nor the subjects know the identity of the treated and untreated subjects during the course of the experiment. Under a double-blind design, the

---

[77] *Cf. id.* at 38 ("[T]he role of the federal government in funding social experiments, although still dominant, has diminished somewhat over time, whereas that of state governments has grown.").

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

researcher remains blind to each subject's group until the researcher has coded all the outcome variables. Researchers who remain in the dark when coding outcomes cannot shade their coding to favor a particular outcome. Hence, double-blind designs can protect against observer bias.[79] Keeping subjects in the dark as to whether or not they are in the treatment group analogously ensures that their behavior and emotional outlook are not biased by the knowledge of how they are being treated. In medicine, the standard way to implement patient ignorance is with placebo-controlled studies. In a placebo-controlled drug study, for example, all patients would receive pills, but the control group would receive a placebo pill, often a sugar pill.[80]

In randomized tests on laws and public information, it will be harder to keep subjects in the dark about how they are being treated or the fact that they are subjects in an experiment. For example, suppose that one were randomly to select certain workplaces to subject to a more rigorous set of workplace-safety standards to help assess the costs and benefits of higher standards. Businesses would need to know which set of workplace safety standards applied to them. The transparency of this randomization, however, is not as significant a concern here as it is in a medical context. Medical researchers are primarily interested in the impact of a drug independent of any placebo effects. In the policy arena, on the other hand, researchers want to see how knowledge of the law affects people's behavior. Information about whether a workplace is treated becomes part of the treatment, but this knowledge is not inherently bad because the researcher's ultimate question is whether a known legal change will have an impact.

There is, however, another problem. Even when subjects do not know whether they are in the treatment or control group, they will typically know that they are participating in a randomized experiment. This knowledge of participation, by itself, may affect results. The impact of knowing that they are being observed might, for example, make subjects alter their behavior to please (or to displease) the researcher. For example, Henry Landsberger recognized this ef-

---

[79] *See* RONALD A. MCQUEEN & CHRISTINA KNUSSEN, INTRODUCTION TO RESEARCH METHODS AND STATISTICS IN PSYCHOLOGY 42-44 (2006) (discussing how to control for a "researcher effect").

[80] In 1862 Austin Flint conducted the first placebo-controlled experiment when he treated a small number of hospital inmates for rheumatic fever. *See* AUSTIN FLINT, A TREATISE ON THE PRINCIPLES AND PRACTICES OF MEDICINE 1019-20 (4th ed. 1873) (discussing his treatment of cases of articular rheumatism with "palliative measures only"). The control group received what Flint called a "placebo," or "placeboic remedy," of a "very largely diluted" "tincture of quassia." *Id.* at 1020.

fect when analyzing the Hawthorne Experiments, which were conducted near Chicago between 1924 and 1927 to determine the effects that better lighting conditions had on workers' performances.[81] In that case, the researchers found a short-term improvement in worker performance after almost any change in lighting,[82] but productivity soon returned to normal levels.[83] Although there remains some controversy over whether the experimental context did affect productivity in that experiment,[84] the label "Hawthorne effect" is now commonly applied to describe changes in behavior attributable to the knowledge by individuals that they are experimental subjects—rather than to the substance of the experimental manipulation.[85] Similarly, the phrase "John Henry effect" is used to describe changes in behavior in comparison groups whose members recognize that they are not being subjected to experimental manipulations.[86]

In medical randomized trials, Hawthorne effects are a concern because the ethical requirement of informed consent necessitates that subjects be informed about and consent to participation in the randomized trial.[87] In the legal context, however, sometimes knowledge

---

[81] *See generally* HENRY A. LANDSBERGER, HAWTHORNE REVISITED 1-27 (1958) (describing the experiments).

[82] *See* Richard Pearson Gillespie, Manufacturing Knowledge: A History of the Hawthorne Experiments 165 (1985) (unpublished Ph.D. dissertation, University of Pennsylvania) (on file with Van Pelt Library, University of Pennsylvania) ("[W]hile production in all three test groups showed an improvement, it could not be correlated with those periods in which lighting was higher.").

[83] *Id.* at 170.

[84] *Compare* WILLIAM H. WHYTE, JR., THE ORGANIZATION MAN 34 (1956) (noting that "output did shoot ahead where conditions were changed, but so did output shoot ahead where no changes had been made"), *with* Stephen R.G. Jones, *Was There a Hawthorne Effect?*, 98 AM. J. SOC. 451, 467 (1992) (finding "essentially no evidence of Hawthorne effects, either unconditionally or with allowance for direct effects of the experimental variables themselves").

[85] *See* Jones, *supra* note 84, at 452-53 (providing numerous examples of authors discussing Hawthorne effects).

[86] *See* Esther Duflo et al., *Using Randomization in Development Economics Research: A Toolkit* ("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)."), *in* 4 HANDBOOK OF DEVELOPMENTAL ECONOMICS, 3895, 3951 (T. Paul Schultz & John Strauss eds., 2008); *see also* 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*, J. ALCOHOL & DRUG EDUC., Spring 1991, at 87, 99 (providing an alleged example of a John Henry effect).

[87] *But cf.* David A. Braunholtz et al., *Are Randomized Clinical Trials Good for Us (in the Short Term)? Evidence for a "Trial Effect,"* 54 J. CLINICAL EPIDEMIOLOGY 217, 219 (2001)

of a change does not necessitate that subjects know that they are taking part in a randomized study. For example, one could imagine a test of speed limits in which the posted limits on different roads were randomly increased or decreased. The drivers on these roads could be informed of the treatment (i.e., the speed limit on that road) without necessarily knowing that they were participating in a randomized experiment. But there may be other cases in which almost all subjects will know that there is a legal experiment. In an experiment on workplace safety, the businesses subject to the new rules would likely find out the reason for the new rules, and it also seems likely that others in the industry would recognize the experimental context. This awareness could lead business owners to be temporarily more cognizant of workplace safety issues, possibly muting the effects of the higher standards. In addition, businesses may act in a particular way or report misleading data because they hope to influence the ultimate policy result. This result is less likely to be a concern, however, if there are a large number of businesses in the experiment, so that each business is likely to have only a small effect.

2. Generalizability

The prior subsection noted the difficulty of extrapolating from a sample with certain informational attributes—such as subjects knowing that they were participating in an experiment—to a population with different informational attributes. There are analogous problems of extrapolation when the tested sample may be unrepresentative of the larger population. James Heckman, with a number of different coauthors, has written extensively about these dangers of "randomization bias" in policy experiments—dangers that "cause[] the type of persons participating in a program [treatment group] to differ from the type that would participate in the program as it normally operates."[88] These dangers may occur as a result of self-selection because

---

(noting that once patients are "exposed to the informed consent process," the Hawthorne effect becomes less relevant).

[88] James J. Heckman & Jeffrey A. Smith, *Assessing the Case for Social Experiments*, J. ECON. PERSP., Spring 1995, at 85, 99; *see also* James J. Heckman & Richard Robb, Jr., *Alternative Methods for Evaluating the Impact of Interventions* (considering the "problem of estimating the impact of interventions in the presence of selection decisions by agents"), *in* LONGITUDINAL ANALYSIS OF LABOR MARKET DATA 156, 158 (James J. Heckman & Burton Singer eds., 1985); James J. Heckman & Jeffrey Smith, *Assessing the Case for Randomized Evaluation of Social Programs* (describing the "selection problem" as the fact that "persons who participate in a program are different from persons who do not participate in the sense that the mean outcomes of participants in the non-

volunteers for an experiment differ from those to whom a policy might apply, or because of what we call "experimenter selection," where the experimental design differs from how a permanent policy would operate in terms of the group affected or in some other way.

a. *Self-Selection*

One problem is that it may be inappropriate to extrapolate from subjects who have volunteered, or at least consented, to be tested to a population containing people who would not volunteer or consent. If the attributes of people that lead them not to consent also lead them to react differently to the treatment, then the treatment may produce different effects on the general population. Volunteers are a self-selecting group of individuals who seek exposure to an experimental policy. The causal impact of the experimental policy on this self-selecting group may be different from the causal impact of the policy on the average individual it affects.[89] Chemotherapy drugs, for example, increase the life expectancy of some cancer patients but decrease the life expectancy of those free of cancer (because of the drugs' side effects).[90] Volunteers for experiments involving chemotherapy drugs may not provide good estimates for the effect of the experimental chemotherapy on cancer patients. Volunteers may generally be sicker than the average cancer patient and therefore ready to try unproven therapies.

The same is true with regard to policies. The volunteers for a policy experiment provide an accurate estimate of the causal effects of the policy only if the volunteers are representative of the group of in-

---

participation state would be different from those of non-participants"), *in* MEASURING LABOUR MARKET MEASURES 35, 45-46 (Karsten Jensen & Per Kongshoj Madsen eds., 1993); James J. Heckman & V. Joseph Hotz, *Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training*, 84 J. AM. STAT. ASSOC. 862, 874 (1989) ("[S]imple specification tests eliminate the most unreliable and misleading estimators that give rise to the sensitivity problem . . . ."). *See generally* James J. Heckman, *Randomization and Social Program Evaluation* (discussing the "benefits and limitations of randomized social experiments" (emphasis omitted)), *in* EVALUATING WELFARE AND TRAINING PROGRAMS 201 (Charles F. Manski & Irwin Garfinkel eds., 1992).

[89] When causal impacts of a treatment vary across individuals, the treatment effect is called "heterogeneous." For a discussion of heterogeneous treatment effects, see James J. Heckman, *Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture*, 109 J. POL. ECON. 673, 690-704 (2001).

[90] *See, e.g.*, Laurence A. Cole et al., Letter to the Editor, *False-Positive hCG Assay Results Leading to Unnecessary Surgery and Chemotherapy and Needless Occurrences of Diabetes and Coma*, 45 CLINICAL CHEMISTRY 313, 314 (1999) (noting that false positives can produce negative results, such as unnecessary chemotherapy, which put one patient into a coma).

dividuals who will be affected by the fully enacted policy. Consider an experimental job-skills program. People who volunteer for such a program may be especially likely to benefit from participation. Experimenters can attempt to control for differential effects, but some of the variables that affect the volunteers' responses to the job-skills program will be unobservable. Volunteers may be particularly disciplined in following the program (raising the impact of the program), and the discipline of volunteers may be unobservable or uncorrelated with other observables.

In this case, the estimated effect of the program for volunteers will be higher than the effect for the average low-skill person, and experimenters cannot adjust their effect estimates to account for discipline. If policymakers consider making the program mandatory for people of a certain skill level, then the experimental estimate of the program's effect using volunteers will be biased. Volunteer experiments can, however, guide policymakers in determining whether to offer—but not mandate—a policy to the general population. Under voluntary programs, the government's offer is in some sense the treatment.

Treatment will sometimes affect volunteers and compelled individuals similarly. In medicine, it is routine to move from randomized tests on volunteers to quasi-mandatory, across-the-board treatment proposals for individuals whose condition is similar to those who were subject to experiments. The problem, however, may generally be more severe for legal experiments, because it may be more difficult in a legal context to control for other characteristics. Perhaps researchers can make some headway in measuring the severity of cancer according to test results and symptoms.[91] But individual psychology and business strategies are so diverse that it will often be difficult to come up with metrics that serve as effective controls.

The government can respond to this "voluntariness" problem by designing tests with mandatory participation. Ethical rules require that patients consent to participation in medical experiments, but the government can, and has, applied different rules and regulations to different individuals and businesses. For example, the Emergency Unemployment Compensation Act of 1991 authorized the U.S. Department of Labor to test the impact of a job-search assistance pro-

[91] *See, e.g.*, Nicholas Wade, *Speed Reading of DNA May Help Cancer Treatment*, N.Y. TIMES, Mar. 9, 2010, at D4 ("Researchers at Johns Hopkins University have developed a way to monitor the progress of a patient's cancer treatment using a new technique for rapidly sequencing, or decoding, large amounts of DNA.").

gram by randomly requiring certain recipients of unemployment insurance to participate in the program.[92]

b. *Experimenter Selection*

Even when an experimenter can compel participation, there is still a danger that 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 a researcher conducts an experiment 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.[93] At times, skepticism about inferences from an experiment on a nonrepresentative population may be justified.[94] For example, a randomized workplace-safety experiment on a sample of small firms might not extrapolate easily to a sample of large firms. It may be feasible sometimes to conduct randomization at a national level, for example in choosing Medicare reci-

---

[92] *See* Emergency Unemployment Compensation Act of 1991, Pub. L. No. 102-164, § 201(c)(3)(A), 105 Stat. 1049, 1056 (providing for the "random selection of eligible individuals for participation in the program and for inclusion in a control group"). This program was in effect from November 17, 1991, to February 5, 1994. *See* I.R.C. § 3304 note; *see also* PAUL T. DECKER ET AL., MATHEMATICAL POLICY RESEARCH, INC., ASSISTING UNEMPLOYMENT INSURANCE CLAIMANTS: THE LONG-TERM IMPACTS OF THE JOB SEARCH ASSISTANCE DEMONSTRATION 7 (2000), *available at* http://www.workforcesecurity.doleta.gov/dmstree/op/op2k/op_02-00.pdf (discussing the program); Marcus Stanley et al., Developing Skills: What We Know About the Impacts of American Employment and Training Programs on Employment, Earnings, and Educational Outcomes 40-42 (Oct. 1998) (unpublished manuscript), *available at* http://www.economics.harvard.edu/faculty/katz/files/stanley_katz_krueger_98.pdf (explaining that various programs in Minnesota, Nevada, New Jersey, South Carolina, and Washington were "conducted as random assignment experiments, making their results particularly reliable").

[93] For a discussion of the transition from local demonstration projects to projects on a national scale, see Duflo, *supra* note 64, at 342-55.

[94] *See, e.g.*, GREENBERG ET AL., *supra* note 4, 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." (footnote omitted)).

pients who will receive extra follow-up phone calls.[95] If policy is to be implemented at a national level, then this sort of experimentation will provide a sound assessment of policy. Often, however, coordination and data-gathering needs may make this process difficult.

Moreover, even if policy is to be implemented at a national level, such a characteristic does not necessarily mean that a single uniform policy will be optimal. Randomized results give powerful and transparent information about the average impact of the law on policy outcomes, but teasing out causal information on subgroups of the population is much more difficult.[96] For example, imagine that a speed-limit study randomizing across different cities shows that twenty mile-per-hour limits produce *more* accidents than thirty mile-per-hour limits. It might still be that small, rural cities fare better with the lower limit. One can run regressions on the results of randomized studies to test whether the average result holds true for subgroups within the tested sample. As long as the treatment is randomly applied across small cities, for example, the small-cities subsample can be viewed as a subexperiment. But because a population can be divided in any number of ways and statistically significant results are likely to exist by chance for some subsamples, researchers will occasionally need to draw admittedly arbitrary lines and must use theoretical considerations to help assess whether statistically significant results for subpopulations seem plausible.

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.[97] Similar problems can affect randomized experiments. Suppose, for example, that the federal government tested the effects of mini-

---

[95] *See supra* notes 8-9 and accompanying text (discussing a Medicare statute requiring randomized testing of programs).

[96] *See* James J. Heckman, *Detecting Discrimination*, J. ECON. PERSPECTIVES, Spring 1998, at 101, 101-03 (discussing the difficulty in "[e]stimating the extent and degree of discrimination" in the labor market).

[97] *See* Duflo, *supra* note 78, at 21 ("Economists are often suspicious of lab experiments, because it is not clear that behavior observed in the lab would still apply when people make 'real' decisions . . . ."). This challenge to effective experiments helps to explain why researchers studying social norms through ultimatum games have experimented in relatively poor countries, where it is feasible to make the stakes large enough to affect experimental subjects' welfare. *Cf.* Robert Slonim & Alvin E. Roth, *Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic*, 66 ECONOMETRICA 569, 570-71 (1998) (reporting on an experiment in the Slovak Republic that had varying stakes to "increase the power of the experiment to detect differences in behavior due to differences in stakes").

mum-wage laws by randomly selecting one percent of adults and 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 practices or risk dissension due to inconsistent wages in order to have the chance to hire a few more workers at a lower wage. Data from such a study, therefore, might not reliably reflect the effect of eliminating the minimum wage.[98]

In addition, 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 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.[99] There is no guarantee that marginal effects will correctly identify the approximate impact of the policy. For example, in the hypothetical concealed-carry experiment,[100] a permanent law might encourage more people to possess concealed handguns than a temporary law, but it is unclear 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 include more law-abiding individuals, but this conclusion is only speculation.

At other times, a temporary law may be a poor proxy for long-term effects because the law will have both dynamic and static effects. For example, studies that seek to assess school-choice plans may fail to capture what proponents of such plans claim is a principal benefit—that free enterprise allows educational entrepreneurs to learn what works over time.[101] Other possible arguments suggest that a static analysis might overestimate the benefits of free choice. In the short term, private schools might be willing to lose money in the hopes of

---

[98] This hypothetical study, however, is not without value. The fact that the change is "not worth the trouble" suggests that the benefits of the experimental policy are simply limited to some degree.

[99] *See, e.g.*, Dean Karlan & Jonathan Zinman, Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts 4 (June 25, 2007) (unpublished manuscript), *available at* www.hks.harvard.edu/inequality/Seminar/Papers/Karlan_071.pdf (reporting on an experiment in which loan officers were "randomly encouraged . . . to approve some marginal applications").

[100] *See supra* note 62 and accompanying text.

[101] *See, e.g.*, Terry M. Moe, *Beyond the Free Market: The Structure of School Choice*, 2008 BYU L. REV. 557, 571 (noting that current voucher programs are "too small and too new" to determine "whether the voucher amounts are large enough to . . . stimulate a sufficient supply-side response by schools over time").

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 argue that with a longer-term study, customers would buy appliances that would help them adjust their electricity use based on time of day.[102]

3. Imperfect Randomization

The above subsections have addressed the danger that the tested population might differ systematically from the more general population applicable to the legal experiment. There is also another problem with randomization—ensuring that the treatment group receives the treatment and the control group does not. A computer can randomize between treatment and control groups, but it is not always easy to ensure that the computer's decisions are followed or that the results are properly measured. Dangers include attrition (randomized individuals or entities dropping out of a study), crossover (control group members receiving the treatment, or vice versa), and spillovers (treatment having some effect on the control group as well).

a. *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 treatment. Consider, for example, studies assessing school improvements in a developing country. A school's random placement into a control group might increase student drop-out rates.[103] If the dropouts tend to be the weaker students, and if the measurement of school success depends on the test results of current students, then attrition produces an artificial hurdle for the treatment. Attrition can also bias results when randomization occurs at the individual level. In a medical study, for example, people who receive the treatment but then suffer severe side ef-

---

[102] *See, e.g.*, Aigner, *supra* note 11, at 46 (noting that "these experiments only allow us to estimate short-run elasticities of demand, given existing appliance stocks") .

[103] *See, e.g.*, Abhijit V. *Banerjee* et al., *Remedying Education: Evidence from Two Randomized Experiments in India*, 122 Q.J. ECON. 1235, 1245 (2007) (noting that in an educational experiment using randomization, "[d]ifferential attrition between the treatment and comparison groups could potentially bias the results").

fects might refuse to participate further in the study, making those who continue with the treatment a nonrepresentative sample.[104]

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.[105] But this solution is imperfect because there may be some unmeasurable difference between those who continue in an experiment and those who drop out. Ultimately, sound statistical judgment is needed to assess the reliability of such models.

A more objective solution is to use matched samples.[106] If someone from the treatment group drops out, a result from the corresponding match in the control group should not be counted either.[107] This approach also can be used when randomization is at the institutional or jurisdictional level, if individuals can be matched across institutions or jurisdictions.[108] With matching, it is not necessary to construct a model ex post that seeks to correct for attrition bias—a method that would increase the danger of subjectivity, which in turn increases the risk that the researcher will be able to choose a specification that meets the researcher's goals. Statisticians would need to assign the original matches based on observable characteristics, but the

---

[104] The converse may also be true. As Banerjee and his coauthors hypothesize, "[I]f weak children were less likely to drop out when they benefitted from a [tutor], this could bias the program effect downwards." *Id.*

[105] *See, e.g.*, Richard B. Miller & David W. Wright, *Detecting and Correcting Attrition Bias in Longitudinal Family Research*, 57 J. MARRIAGE & FAM. 921, 922 (1995) (describing the standard method of responding to this problem by incorporating a variable representing the probability of dropping out directly into the study (citing James J. Heckman, *Sample Selection Bias*, 47 ECONOMETRICA 153 (1979) and James J. Heckman, *The Common Structure of Statistical Methods of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models*, 5 ANNALS ECON. & SOC. MEASUREMENT 475 (1976)).

[106] *See* Duflo et al., *supra* note 86, at 3925 ("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.").

[107] So, for example, if an experiment were to test whether eating at a government-run restaurant that served only healthy food improved health outcomes, subjects would be randomized into the group that received free meals at the restaurant and a control group. For each person in the treatment group who decided to stop coming to the restaurant, the matched person in the control group would also not be counted.

[108] Continuing the example from the previous footnote, *supra* note 107, if some jurisdictions were randomly selected and offered free federally provided healthy food restaurants but then declined them, results from the matching control jurisdictions also would not be counted.

matches would be difficult to manipulate because the researchers will not know in advance who will drop out.

b. *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, it is not easy to escape. Nonetheless, imperfections may occur, especially if the government wishes to leave some room for later discretion. For example, crossover can 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 reassign their children from larger to smaller classes.[109] This sort of influence dilutes the treatment, as the smaller classes become larger, and it also means that the treated students on average will come from families with relatively high motivation.[110]

Once again, statistical correctives exist. Under an "intent-to-treat" methodology, an individual or entity who crosses over is counted with the group to which that person or entity was originally assigned.[111] This practice reduces the measured effect of the treatment. Statisticians, however, can compensate for the bias that the intent-to-treat approach introduces with a simple mathematical formula.[112] Assuming that it is possible to measure who ended up receiving the treatment and who received the control, the formula can be applied mechanically, without allowing the experimenters any discretion. This correction will generally improve the estimate of the treatment effect.[113] This ad-

---

[109] *See* Alan B. Krueger, *Experimental Estimates of Education Production Functions*, 114 Q.J. ECON. 497, 505 (1999) (reporting lower attrition rates of students in smaller classes, with some exceptions).

[110] *See id.* at 506 ("[I]f the movement between class types was associated with student characteristics (e.g., students with stronger academic backgrounds more likely to move into small classes), these transitions would bias a simple comparison of outcomes across class types.").

[111] *See, e.g.*, Guido W. Imbens & Joshua D. Angrist, *Identification and Estimation of Local Average Treatment Effects*, 62 ECONOMETRICA 467, 472 (1994) (discussing the intent-to-treat approach).

[112] 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 64, at 354 (citing Imbens & Angrist, *supra* note 111).

[113] *See* Imbens & Angrist, *supra* note 111, at 470 (describing a theorem that "implies that local average treatment effects can be estimated by comparing the average of outcome Y and treatment D at two different values of the instrument Z").

justment is imperfect, though, because those who cross over may differ systematically from those who do not.[114] Once again, this flaw illustrates that even with randomized statistical methodologies, such statistical judgment may be needed to interpret the study results.

c. *Spillovers*

The final danger of randomization is that the treatment will spill over onto the control group. Suppose, for example, that a shame sanction reduces recidivism not only among those who are shamed, but also among those who are randomized to the control group and happen to 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 do. The comparison of treatment and control groups will underestimate the effects of the intervention because the control group has adopted the treatment as well. 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 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.

Sometimes, a feasible solution is to randomize across geographical areas, rather than across individuals. Edward Miguel and Michael Kremer showed, for example, that randomized studies at an individual level underestimated the benefits of deworming drugs, which benefited those in the immediate area who had not received the drugs.[115] Randomizing across geographical areas largely solved the problem.[116] 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.[117]

---

[114] *See id.* at 472 (providing examples of when the application of the formula would be and would not be appropriate).

[115] Edward Miguel & Michael Kremer, *Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities*, 72 ECONOMETRICA 159, 208 (2004).

[116] *Id.*

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

B. *Other Issues*

1. Costs

Experimental costs include implementation costs and direct experimental-policy costs. Other things equal, the lower these costs for a given policy, the stronger the argument for experimentation. Implementing a policy experiment can be expensive. Policymakers must first overcome obstacles to experimentation, such as citizen opposition. When opposition to randomization is high, convincing the experimental subjects that the experiment is in their interest may necessitate more effort than the value of the information that the experiment would yield. Once an experiment is approved, the implementation costs continue. Policymakers must inform individuals subject to the experimental policy about the change in policy while making clear to the rest of the population (the control group) that their policy landscape remains unchanged. Adding to the complexity, a policy experiment's "laboratory" is the everyday world of human behavior, rather than the controlled setting of the scientific lab.

This landscape creates several complications. First, policymakers must determine means to measure the outcomes of interest. At times, preexisting data-collection efforts may reflect the outcomes of interest, but, at other times, new sources of data on outcomes must be found. Such data-gathering efforts will be costly. Second, policymakers must confront the noncompliance problem. Individuals are not mice and may find ways to avoid complying with the experimental treatment. Policymakers must find legitimate means of limiting such noncompliance, but such means will generally be costly. Even so, there will always be some number of noncompliers, and policymakers must ascertain means of preventing attrition and noncompliance from biasing the results of the experiment.

It is possible to use randomized experimentation to test many different variations on policies. For example, a test of speed limits could allow for a wide range of speed limits, or it could test whether a tailored policy (of thirty miles-per-hour in small cities and twenty miles-per-hour in large cities) is superior to an untailored policy (of thirty miles-per-hour in all cities).[118] But the possibilities for tailoring

[118] Randomized testing of this kind on the Internet has shown, for example, that tailoring a retail website's landing pages to depend on specific search queries produces more sales than does a one-size-fits-all homepage. For example, clicking on a Google advertisement for http://www.musiciansfriend.com after searching for "electric guitar" will take you to a different page than will clicking on the same advertisement after

in any particular arena are endless, and it is unreasonable to expect that more than a tiny fraction of these possibilities will ever be tested. Hence, it will be important for lawmakers and regulators to use theory and intuition to guide their choice of scarce options in order to determine with full awareness whether untested policies may still be preferable to tested polices. Sometimes, it may be worthwhile to focus on what seems to be the most attractive possibility, even if there is some chance that a more attractive option will emerge later.

Happily, the costs of experimental design and implementation are subject to economies of scale. If legislators and administrators begin to implement many experiments, then they will learn effective techniques for experimentation. In addition, public familiarity with experimental processes may reduce the costs of convincing the public to participate in experiments and may reduce the costs of explaining the experimental policy to the subjects of the policy. Therefore, the marginal costs of experimental policies should decline with the number of policies. A widespread and systematic emphasis on policy experimentation would decrease costs with respect to the current practice of piecemeal government-policy experimentation.

Economies of scale, however, reduce the marginal costs of experimentation but cannot eliminate them. As a result, policymakers should first experiment with policies that have low experimentation costs, all else being equal. While it is impossible to describe completely the factors influencing the costs of experimentation, several salient policy features are worth examining. Most obviously, policymakers should experiment with policies that have relatively positive expected effects.[119] In other words, policymakers should experiment with the best candidates first. This strategy will reduce the direct costs of experimentation on the subjects of the experimental policy. Meanwhile,

---

searching for "electric bass" because randomized testing of contingent strategy by Omniture showed a higher revenue of fifteen percent per customer when the landing pages were tailored to the specific search queries. Conversation between Ian Ayres and Matt Roche, President, Omniture (June 14, 2007); *see also* AYRES, *supra* note 14, at 55-56 (describing Google AdWords); Paat Rusmevichientong & David Williamson, *An Adaptive Algorithm for Selecting Profitable Keywords for Search-Based Advertising Services* (same), *in* PROCEEDINGS OF THE 7TH ACM CONFERENCE ON ELECTRONIC COMMERCE 260, 260 (Joan Feigenbaum et al. eds., 2006). For more information on Google AdWords, see generally Wikipedia, http://en.wikipedia.org/wiki/AdWords (last visited Jan. 15, 2011), and text accompanying *infra* note 192.

[119] *See* Yair Listokin, *Learning Through Policy Variation*, 118 YALE L.J. 480, 513-14 (2008) (arguing that, in many cases, the expected effect of policy is less important than the variance of the expected effects, but other things being equal, higher-expected value policies are superior to lower-expected value policies).

experiments should generally be as modest as possible, but still big enough to have measurable effects.

An additional consideration is that concentrated populations of experimental subjects are likely to have lower experimental implementation costs than will diffuse subject populations. Informing the entire national population of the existence of a randomized experiment and of each individual's status as subject or control within the experiment is likely to be prohibitively expensive. By contrast, informing each company on the New York Stock Exchange (NYSE) of the existence of an experiment, as well as the company's experimental status, will be much easier. The population of NYSE companies is clearly defined and finite, reducing the costs of the experiment. As a result, policymakers should first pursue experimental policies when the target population of the policy is small, *ceteris paribus*.

2. 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 resolving whether the government must obtain informed consent in such experiments. 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,[120] where any participant may choose not to receive the services that the government offers.[121] Second, an existing collection of essays already explores this issue in considerable detail.[122]

This subsection will address ethical concerns, summarizing and developing 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 experimentation

---

[120] *See supra* notes 8-9 and accompanying text (describing the program requiring nurses to call Medicare patients in an attempt to reduce costs).

[121] There may, however, be objections based on inequality among those who volunteer for the experiment, an issue to which the Article will return below. *See infra* subsection III.B.3 (discussing equality concerns for individuals assigned to less desirable experimental groups).

[122] *See* Alice M. Rivlin & P. Michael Timpane, *Introduction and Summary* to ETHICAL AND LEGAL ISSUES OF SOCIAL EXPERIMENTATION 1, 1 (Alice M. Rivlin & P. Michael Timpane eds., 1975) (introducing a collection of papers discussing the "new ethical and legal issues that need to be carefully examined" when testing new social policies).

from medical experimentation,[123] where informed consent is generally required.[124] Medical treatment to which a patient does not consent violates the patient's bodily-integrity rights;[125] the problem is not randomization. Similarly, 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. For example, if a person has a right not to have property taken by the state,[126] then the state cannot take that property in an experiment. But to the extent that a government would be authorized to enact a policy generally, on the Lockean theory of implicit consent,[127] there should be no ethical bar to the state's enacting the policy against only 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.[128]

---

[123] 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.

Id. at 5.

[124] *See* Kathryn A. Tuthill, Commentary, *Human Experimentation: Protecting Patient Autonomy Through Informed Consent*, 18 J. LEGAL MED. 221, 221 (1997) ("The doctrine of informed consent requires that a physician inform a patient or research subject of the benefits, risks, and alternatives to medical treatment or experimental procedures before such treatment is administered.").

[125] 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 in damages." Schloendorff v. Soc'y of N.Y. Hosp., 105 N.E. 92, 93 (N.Y. 1914), *abrogated on other grounds by* Bing v. Thunig, 143 N.E.2d 3 (N.Y. 1957).

[126] *See* U.S. CONST. amend. V (requiring that private property not "be taken for public use, without just compensation").

[127] *See* Peter G. Brown, *Informed Consent in Social Experimentation: Some Cautionary Notes* ("'[E]very man that hath any possession or enjoyment of any part of the dominions of any government doth thereby give his tacit consent . . . .'" (quoting John Locke, *An Essay Concerning the True Original, Extent and End of Civil Government*, *in* THE ENGLISH PHILOSOPHERS FROM BACON TO MILL 403, 452 (Edwin A. Burtt ed., 1939))), *in* ETHICAL AND LEGAL ISSUES OF SOCIAL EXPERIMENTATION, *supra* note 122, at 79, 96.

[128] *See* IMMANUEL KANT, GROUNDWORK OF THE METAPHYSIC OF MORALS 96 (H.J. Paton trans., Harper Torchbooks 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." (emphasis omitted)).

This principle, however, 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—because policymakers also believe it has enough of a chance of success to make it worth trying. If this action counts as treating people as a means only, then the ethical permissibility of a new policy must be judged without consideration of any informational benefits that implementation of the policy might produce. But many legal regimes, such as patent law and securities law, are justified in part on the basis that they improve information.[129] Information produced by a policy about the policy itself should not be condemned as irrelevant.

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.[130] Although this conclusion 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."[131] He also notes that the Nuremberg principles on medical experimentation emphasized the importance of experimental design.[132] Thus, if universal experimentation is permissible, there is an a fortiori argument that random experimentation must be permissible as well. The difference between the universal experiment and the random experiment is that some people do *not* receive the treatment. Unless there is an equality right to receive the treatment,[133] this difference should not make the experiment more problematic.

Medical experimentation itself further supports the argument that if universal policy experiments are permissible, then randomized policy experiments must be permissible as well, because medical experiments can generally be viewed as equivalents to policy experiments. Subjects in medical experiments who give informed consent presumably would prefer a guarantee of receiving the treatment rather than a chance of receiving a placebo. The status quo is a legal regime that constrains liberty by forbidding distribution of the treatment. Let us

---

[129] *See generally, e.g.*, Jeanne C. Fromer, *Patent Disclosure*, 94 IOWA L. REV. 539 (2009) (discussing the patent system's goal of encouraging disclosure of inventions).

[130] *See supra* Part I.

[131] Robert M. Veatch, *Ethical Principles in Medical Experimentation*, *in* ETHICAL AND LEGAL ISSUES OF SOCIAL EXPERIMENTATION, *supra* note 122, at 21, 37.

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

[133] *See infra* subsection III.B.3.

assume that the legal prohibition on what Eugene Volokh has called "medical self-defense" is permissible.[134] When the government authorizes a medical experiment,[135] it is effectively authorizing a new legal regime that permits patients to have access to a treatment. The government, however, does not authorize this new legal regime in a universally applicable way; instead, it 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 system suggests that policy randomization is permissible, at least so long as the group being randomized gives informed consent. The argument for informed consent, however, 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 be subject to Policy *Y* only 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 subjects in a legal experiment, in which they might be denied the right to the medicine at random (the equivalent to a medical experiment in which they might receive a placebo instead of the treatment).

Those who defend the legitimacy of medical experimentation must either develop an account by which baselines are permissible or allow legal policymakers to play the same game outside the medical context. Existing randomized legal experiments generally allow subjects to opt into an apparently more favorable treatment. For example, drug offenders may receive the option to participate in an experiment in which they might be randomly assigned to a drug court.[136]

---

[134] Eugene Volokh, Essay, *Medical Self-Defense, Prohibited Experimental Therapies, and Payment for Organs*, 120 HARV. L. REV. 1813, 1815 (2007); *see also* Abigail Alliance for Better Access to Developmental Drugs v. Von Eschenbach, 495 F.3d 695, 713 (D.C. Cir. 2007) (en banc) (finding no constitutional right to have access to "investigational drugs").

[135] Governmental involvement 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 N.Y.U. J. LEGIS. & PUB. POL'Y 295, 305 (2000) (describing Phase II as "the first stage of testing at which drug efficacy becomes a formal consideration").

[136] *See, e.g.*, Denise C. Gottfredson & M. Lyn Exum, Research Note, *The Baltimore City Drug Treatment Court: One-Year Results From a Randomized Study*, 39 J. RES. CRIME & DELINQ. 337, 343 (2002) (detailing a program in which 235 eligible clients were randomly assigned either to drug court or to standard criminal processes).

This characteristic makes experimentation a one-way ratchet, allowing testing of a more lenient alternative within an existing draconian regime, but not allowing testing of more draconian legal approaches. The only way to test the more draconian approaches would be to change the baseline to those approaches and then allow individuals to opt into an experiment in which they might receive more lenient treatment. Similarly, policymakers could test raising the speed limit by allowing drivers to opt into a program in which they are permitted to drive ten miles-per-hour over the limit, but in order to test lowering the speed limit, policymakers would have to change the baseline. An insistence on informed-consent privileges the status quo legal regime over alternatives—even if a universal application of neither the status quo nor the alternative violates any rights.

3. Equality Concerns

Concerns about informed consent focus on the rights of those subject to the experiment. Concerns about equality, on the other hand, focus on the rights of those who are either 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; it extends also to cases in which a government with limited resources distributes those resources at random.[137] For example, governments have used lotteries to distribute scarce low-income housing,[138] rights to immigrate,[139] and positions in magnet and charter schools.[140] Maurice Rosenberg points out that random experimentation may inevitably be 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

---

[137] Such distribution has generally raised fewer objections than randomization for experimental purposes alone, and, as a result, experimentation has been particularly feasible in cases where arbitrary decisions needed to be made in any eventuality. *See* GREENBERG ET AL., *supra* note 4, 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").

[138] *See, e.g.*, Denise Irene Arnold, *Lottery Prize Is Affordable Homes*, N.Y. TIMES, Feb. 7, 1988, § 21 (Long Island Weekly ed.), at 12 (discussing a local housing lottery on Long Island).

[139] *See, e.g.*, 8 U.S.C. § 1153(e)(2) (2006) (providing for distribution of visas for diversity immigrants "strictly in a random order").

[140] *See, e.g.*, Cullen et al., *supra* note 32 (analyzing such a lottery in the Chicago public schools).

the differences in treatment."[141] If equal protection is 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 of these contexts. In the experimental context, randomization has benefits that have already been discussed,[142] 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.[143] In the United States, the equal protection 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.[144] In the leading case on this issue, 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." [145] A court someday might fail to follow or even overturn this precedent, but for now the precedent reinforces the plausibility of the legal argument that randomization does not violate the Equal Protection Clause.[146]

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."[147] "Why should Parliament," he asks,

---

[141] Maurice Rosenberg, *The Impact of Procedure-Impact Studies in the Administration of Justice*, LAW & CONTEMP. PROBS., Summer 1988, at 13, 16.

[142] *See supra* Part II.

[143] Rent-seeking can still occur if large numbers of individuals spend money to enter the lottery. *See, e.g.*, Thomas W. Hazlett & Robert J. Michaels, *The Cost of Rent-Seeking: Evidence from Cellular Telephone License Lotteries*, 59 S. ECON. J. 425, 425 (1993) (analyzing government lotteries for cellular-telephone licenses that led to over 320,000 applications).

[144] For a landmark case explaining rational basis review under the Equal Protection Clause, see *Williamson v. Lee Optical Co.*, 348 U.S. 483 (1955).

[145] Aguayo v. Richardson, 473 F.2d 1090, 1109-10 (2d Cir. 1973). One commentator has criticized the court for not indicating that its decision would be valid only for as long as the "value of the program remained uncertain." Note, *Reforming the One Step at a Time Justification in Equal Protection Cases*, 90 YALE L.J. 1777, 1783 (1981).

[146] Randomization schemes, however, may sometimes violate other constitutional provisions. *See, e.g.*, Delaware v. Prouse, 440 U.S. 648, 663 (1979) (finding random stops of vehicles to check drivers' licenses and registrations inconsistent with the Fourth Amendment).

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

"not make abortion criminal for pregnant women who were born in even years but not for those born in odd ones?"[148] 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.[149] Dworkin claims that checkerboard statutes offend a principle that he calls "integrity."[150] 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."[151] That requirement does not occur with random experimentation, where a single principle—the need to obtain more information—justifies both the treatment and control conditions.[152]

Dworkin's concern is that randomness seems arbitrary, but arbitrariness is often more troubling when it is nonrandom. Consider, for example, the different approaches of Justice Stewart and Justice Marshall in *Furman v. Georgia*[153] to the question of whether the death penalty is so capricious as to deny due process. Justice Stewart criticized a state's criminal system because "of all the people convicted of [capital crimes], many just as reprehensible as these, the petitioners [in *Furman* were] among a capriciously selected random handful upon whom the sentence of death has in fact been imposed."[154] Justice Marshall, meanwhile, observed that "[i]t also is evident that the burden of capital punishment falls upon the poor, the ignorant, and the underprivileged members of society."[155] If Justice Marshall was correct (and there is abundant evidence that he was)[156] in claiming that the death

---

[148] *Id.* at 178.

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

[150] *Id.* at 183-84.

[151] *Id.* at 184.

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

[153] 408 U.S. 238 (1972) (per curiam).

[154] *Id.* at 309-10 (Stewart, J., concurring) (footnote omitted); *see also id.* at 293 (Brennan, J., concurring) ("[I]t smacks of little more than a lottery system."); *id.* at 309 (Stewart, J., concurring) ("These death sentences are cruel and unusual in the same way that being struck by lightning is cruel and unusual."); *id.* at 313 (White, J., concurring) ("[T]here is no meaningful basis for distinguishing the few cases in which it is imposed from the many cases in which it is not.").

[155] *Id.* at 365-66 (Marshall, J., concurring).

[156] *See* DAVID C. BALDUS ET AL., EQUAL JUSTICE AND THE DEATH PENALTY 133 (1990) (concluding from the data in their empirical study that "excessive sentences

penalty is disproportionately visited upon the "poor, the ignorant, and the underprivileged," then Justice Stewart cannot be right in alleging that the death sentence is randomly assigned. Justice Marshall's concern resonates with ex ante equal protection concerns because citizens are treated differently from the very beginning as a result of arbitrary characteristics. Justice Stewart's concern instead resonates with an ex post equal protection perspective. Truly random application of law provides each citizen with ex ante equality—an equal chance of being assigned to the same legal rules. A constitutional or moral concern with truly random application of law instead turns on arbitrarily treating equal people differently ex post.

Many observers of the legal system may have a more visceral, negative reaction to ex-post randomness than to ex-ante randomness. Justice O'Connor, in *Ohio Adult Parole Authority v. Woodard*, expressed a concern with a hypothetical clemency procedure: "Judicial intervention might, for example, be warranted in the face of a scheme whereby a state official flipped a coin to determine whether to grant clemency . . . ."[157] This kind of language suggests that courts might be hostile to a truly random application of the law. The New York State Commission on Judicial Conduct in 1982 served a complaint on Alan I. Friess, a Manhattan Criminal Court judge for, among other things, deciding in open court between a twenty- and thirty-day criminal sentence by flipping a coin.[158] More recently, the Virginia Supreme Court similarly removed trial judge James Michael Shull from office for, among other things, determining custody rights for a Christmas holiday by flipping a coin.[159] The Supreme Court rejected Judge Shull's rationale that the probabilistic decision was an attempt "to encourage the litigants to resolve the custody issues by themselves."[160] Federal Judge Gregory A. Presnell similarly used randomization as "a new form of alternative dispute resolution" when he ordered two attorneys

---

most likely result from suspect and illegitimate factors, such as the race and socioeconomic status of the defendant or the victim, or, perhaps, other idiosyncratic factors").

[157] 523 U.S. 272, 289 (1998) (O'Connor, J., concurring).

[158] *See In re* Friess, 457 N.Y.S.2d 33, 34-35 (1982) (denying severance of the charge against Judge Friess for coin-flipping from a charge that he resolved a different dispute "by submitting it to a show of hands by spectators in the courtroom").

[159] *See* Judicial Inquiry & Review Comm'n v. Shull, 651 S.E.2d 648, 658 (Va. 2007) (removing Judge Shull for several instances of misconduct, including coin-flipping as well as ordering a victim of domestic abuse to show wounds on her thigh in court and making an "improper ex parte telephone call"); *see also* Gary Slapper, *Weird Cases: Justice by Coin-Toss*, TIMES ONLINE (London), Nov. 16, 2007, http://business.timesonline.co.uk/tol/business/law/article2882090.ece (describing the Friess and Shull cases).

[160] *Shull*, 651 S.E.2d at 652.

to resolve a deposition-location dispute by playing a game of rock-paper-scissors.[161]

While many people are viscerally appalled by the notion of judges flipping coins to decide legal issues, coin flipping need not be a meaningless ritual. In particular contexts, there are a variety of public policy rationales for randomized decisions. It is not clear whether Judge Shull was sincere in claiming that his coin flipping over child custody was an attempt to promote private dispute resolution. But the rationale is not implausible. Indeed, one of us has shown that probabilistically dividing an entitlement by randomly giving it to one disputant or another can in fact promote private settlement.[162] Disputants bargaining in the shadow of probabilistically divided, Solomonic rights have powerful incentives to speak more honestly with each other—and they therefore may be more likely to settle a dispute before the actual coin flip,[163] just as the lawyers in the deposition dispute resolved their dispute before having to play rock-paper-scissors on the courthouse steps.[164] Moreover, in the context of child custody, Jon Elster has proffered an independent rationale for resolving custody disputes by coin flipping.[165] Elster argues that probabilistically assigning custody in close cases is valuable because the state does not tell the child that one

---

[161] Adam Liptak, *Lawyers Won't End Squabble, So Judge Turns to Child's Play*, N.Y. TIMES, June 9, 2006, at A19. Liptak further reported that "[c]hildish lawyers are commonplace, but the use of children's games to resolve litigation disputes is apparently a new development." *Id.*; *see also* Jeralyn E. Merritt, *The "Rock, Paper, Scissors" Judge*, TALKLEFT (June 9, 2006, 5:06:39 AM), http://www.talkleft.com/story/2006/06/09/305/45461 (defending Judge Presnell's acumen in the wake of the rock-paper-scissors story).

[162] *See* Ian Ayres & Eric Talley, *Solomonic Bargaining: Dividing a Legal Entitlement to Facilitate Coasean Trade*, 104 YALE L.J. 1027, 1073-78 (1995) (demonstrating that awarding property rights probabilistically is efficient given that all parties have knowledge of the court's probability distribution).

[163] Solomonic entitlements have an "information-forcing" effect on ex ante bargaining because disputants are no longer simply buyers or sellers. In traditional negotiations, sellers overstate their valuations, and buyers understate their valuations, making it difficult to discover all instances of value-enhancing trade. But in the shadow of randomized asset allocation, it is possible for plaintiffs to enter into two different kinds of settlement: one where they buy the defendants' probabilistic entitlements and one where they sell their own probabilistic entitlements. The offsetting incentives to overstate value as a seller and understate value as a buyer lead to more forthright and efficient negotiations. *See id.* at 1045-47 (discussing the information-forcing effect of untailored liability rules); *see also* Peter Cramton et al., *Dissolving A Partnership Efficiently*, 55 ECONOMETRICA 615, 626 (1987) (arguing that a "simple bidding game" can "achieve[]" an "ex post efficient allocation" of assets when dissolving a partnership).

[164] *See* Liptak, *supra* note 161 (noting that the lawyers agreed to meet prior to playing the game).

[165] *See* JON ELSTER, SOLOMONIC JUDGMENTS: STUDIES IN THE LIMITATION OF RATIONALITY 170-72 (1989) (discussing arguments for and against coin-flipping).

parent is marginally better than the other.[166] For Elster, publicly stating that the mother or father is the marginally better Christmas custodian may not be in the best interest of the child.[167]

Judicial antipathy to randomized decisions is at its highest with regard to decisionmaking in criminal cases. But even here, it is not difficult to conjure public policy rationales for coin-flipping sentences. It is elementary economics that probabilistically uncertain sentences will have a greater deterrence effect with regard to risk-averse defendants than will certain sentences.[168] New York State might get a bigger bang for its incarceration buck if it followed Judge Friess and flipped coins for twenty- and thirty-day sentences instead of sentencing everyone to twenty-five days.[169] This deterrence result is, however, reversed for risk-preferring criminals, and it is thus reassuring that Judge Friess, before flipping, asked the defendant if he was a "gambling man."[170] But to our minds, an even stronger rationale for randomization—even with regard to criminal sentencing—is to promote learning. After centuries of experience, we still do not have definitive evidence on whether longer sentences rehabilitate or harden criminals.[171] Justice O'Connor is appalled by the idea of clemency by chance, and randomness applied in a single case seems unlikely to produce useful in-

---

[166] *See id.* at 171 ("One may well imagine a coin-tossing problem coming to symbolize the equal worth of the parents . . . ."); *see also* Adam M. Samaha, *Randomization in Adjudication*, 51 WM. & MARY L. REV. 1, 20-21, 29-30 (2009) (discussing Elster's argument in more detail).

[167] *See* Samaha, *supra* note 166, at 29 (explaining that Judge Brown "faced a choice that other judges might have decided on questionable grounds—for example, by a tie-breaking preference for older couples over fathers, or vice versa").

[168] *See* Steven Shavell, *Economic Analysis of Public Law Enforcement and Criminal Law* 2 (Nat'l Bureau of Econ. Research, Working Paper No. 9698, 2003), *available at* http://www.nber.org/papers/w9698.pdf (arguing that if parties are risk averse, then it may be "socially beneficial" to impose lesser sanctions than if the parties are risk neutral).

[169] *Cf.* David Lewis, *The Punishment That Leaves Something to Chance*, 18 PHIL. & PUB. AFF. 53, 58-62 (1989) (defending punishments where the severity is randomized, en route to justifying harsher penalties for those who, by coincidence, cause more harm); Florynce Kennedy, Letter to the Editor, *For Whom the Coin Is Tossed*, N.Y. TIMES, Feb. 13, 1982, at 24 (comparing Judge Friess's coin toss to other intrinsic uncertainties in the judicial system).

[170] Slapper, *supra* note 159. However, the coin toss's general deterrent effect depends less on whether the particular defendant being sentenced likes to gamble than on whether a prospective criminal is risk preferring or risk averse. *See supra* note 168 and accompanying text.

[171] For an argument that longer sentences have minimal deterrent effect, see John M. Darley, *On the Unlikely Prospect of Reducing Crime Rates by Increasing the Severity of Prison Sentences*, 13 J.L. & POL'Y 189 (2005).

formation.[172] There might, however, be value in randomly granting clemency and parole to inmates selected at random to see if, in fact, they have a higher recidivism rate than do those who are not selected.[173]

The informational rationale for randomization also acts as a principle for deciding when not to test and when to stop testing. We should not allow randomized tests of parachutes[174] because we already have strong evidence that they are effective. And it is standard protocol to shut down medical trials early if it becomes clear that either the control or treatment therapy is superior.[175] The case for randomized testing is at its strongest when the evidence is truly in equipoise about which of two policies is the best. It is analytically convenient to contrast extreme examples of knowledge (as in the parachute example) and ignorance (as in the concept of evidentiary equipoise). But in many cases, existing evidence does not compel the conclusion that either the treatment or the control is more likely to be effective.[176] Indeed, even if we start in a position of evidentiary equipoise, as any randomized trial proceeds, the very process of learning destroys the equipoise and creates the vexing problem of partial information.[177] Notwithstanding the supposed requirements of informed consent, medical trials routinely fail to give participants the best current infor-

---

[172] *See supra* note 157 and accompanying text.

[173] *Cf.* Jeffrey R. Kling et al., *Experimental Analysis of Neighborhood Effects*, 75 ECONOMETRICA 83, 84 (2007) (analyzing a "randomized experiment in which some families living in high-poverty U.S. housing projects were offered . . . housing vouchers . . . while others were not").

[174] *See* Gordon C.S. Smith & Jill P. Pell, *Parachute Use to Prevent Death and Major Trauma Related to Gravitational Challenge: Systematic Review of Randomised Controlled Trials*, 327 BMJ 1459, 1460 (2003) (noting ironically that "[t]he basis for parachute use is purely observational, and its apparent efficacy could potentially be explained by a 'healthy cohort' effect").

[175] *See* Sarah J.L. Edwards et al., *The Ethics of Randomised Controlled Trials from the Perspectives of Patients, the Public, and Health Care Professionals*, 317 BMJ 1209, 1209 (1998) ("The scientific rationale for conducting a trial rests in collective equipoise, which means that the medical community as a whole is genuinely uncertain over which treatment is best.").

[176] Moreover, from an efficiency perspective, it is sometimes cost effective to test and eliminate low-probability therapies that might teach us a great deal. *See* Wallace, *supra* note 53, at 434 (explaining that "unacceptably large variances together with large expenses in obtaining more information in the form of additional sample data dictate the need for restrictions on parameters"); Martin L. Weitzman, *Optimal Search for the Best Alternative*, 47 ECONOMETRICA 641, 649 (1979) ("The purpose of the model formulated in this paper is to sharply characterize [the] optimal search among alternative sources with different characteristics.").

[177] *See* Richard J. Lilford & Jennifer Jackson, *Equipoise and the Ethics of Randomization*, 88 J. ROYAL SOC'Y MED. 552, 554-55 (1995) (discussing the implication of partial information on informed consent).

mation about the likely result of the trial.[178] The reason for the failure is to keep patients participating. Patient surveys indicate, unsurprisingly, that "[w]illingness to undergo randomisation drops as prospective participants are given more preliminary data and as they are made aware of any accumulating evidence of effectiveness."[179]

IV. GUIDELINES AND APPLICATIONS

We saw in Part II that even with the best statistical tools, it is often difficult to make inferences about causality from nonrandomized policy changes. Randomization generally makes interpretation much easier, even though, as we saw in Part III, randomized experiments can be difficult to interpret. Given these concerns, this Part develops some general guidelines for randomized experimentation, describes how legislatures and administrative agencies might initiate randomized studies, and offers some specific applications of randomizing law.

A. *General Guidelines*

In many respects, randomized experiments should conform to ordinary principles of experimentation. For example, the sample should be large enough to generate meaningful results.[180] 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. The more observations, the better the chance that any actual effect will be correctly 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 a threshold for identifying an experiment as a success. Statisticians have long recognized these thresholds as arbitrary.[181]

---

[178] *Cf.* Edwards et al., *supra* note 175, at 1209 ("Most doctors expressed willingness to enter their patients in trials even when the treatments offered were widely available but were not an equal bet prospectively . . . .").

[179] *Id.* Patients might be more willing to accept randomization if they knew that the trial would increase their probability of getting the more effective therapy. But when one therapy is known to be more likely effective, self-interested patients would prefer to receive 100% of that therapy.

[180] *See, e.g.*, Duflo et al., *supra* note 86, at 3918-28 (discussing the issue of sample size in randomized experiments).

[181] *But see* Lester V. Manderscheid, *Significance Levels—0.05, 0.01, or ?*, 47 J. FARM ECON. 1381, 1381 (1965) (urging that the level of statistical significance employed in a particular experiment is "not arbitrary but rather is, or at least should be, a deliberate choice").

Meanwhile, policymakers must consider the unit of analysis at which randomization occurs.[182] 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, to reduce attrition bias, policymakers should generally use matched samples, with matching occurring before the experiment on all available variables.[183]

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, whether to opt out.[184] This practice 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.[185] But governments—at least in theory—could make participation in a randomized experiment mandatory (just as they have done with the draft lotteries), and they could 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.[186]

After accounting for experimental implementation costs, which are fixed, one should find that the threshold for implementing an experiment should be lower than the threshold for enacting new policies. While policies apply to everyone indefinitely, the direct effects of

---

[182] *See, e.g.*, Duflo et al., *supra* note 86, at 3929-30 ("An important practical design is whether to randomize the intervention at the level of the individual, the family, the village, the district, etc.").

[183] *See supra* note 106 and accompanying text (discussing the use of matched samples).

[184] *See generally* Rivlin & Timpane, *supra* note 122 (discussing a wide range of experiments).

[185] *See generally* Tuthill, *supra* note 124, at 221-46 (providing an overview of law concerning informed-consent requirements in medical experimentation).

[186] 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 could simply be ignored in an experiment on fee shifting in court cases. More generally, it is possible to estimate intent-to-treat effects that look at the impact of treatment offers or attempts, regardless of whether the subjects comply. *See supra* notes 111-14 and accompanying text (discussing the "intent-to-treat methodology").

experiments apply to only a subset of the population for a discrete period of time. As a result, the downside of implementing an experimental policy is much lower than the downside of an ordinary policy, implying that the threshold for experimental policy implementation is lower than the threshold for permanent enactment. Moreover, the informational value of an experiment is higher than the informational value of ordinary policy enactment. Experiments allow for better identification of the causal effects of policies than do ordinary policy changes. When the policy environment does not change radically over time, the information from the experiment yields benefits over a long period. Randomized experiments thus provide uniquely accurate information with long-lasting value.

A policy can be randomly assigned at many different levels of randomization. Some policies can be randomly assigned at the individual level. This level of randomization is familiar from the pharmaceutical industry. In a drug trial, some individual subjects are given the experimental drug, while other individuals who serve as controls receive the drug that constitutes the existing state-of-the-art treatment.[187] Similarly, individuals can be randomized into different policies. For example, Medicare's prescription drug program, "Part D," randomly assigned more than six million people to one of up to twenty qualified state plans.[188] Recipients were free to opt out, but the legal default for the individual was chosen at random.

In other cases, randomization may take place at a different level of generality. It makes little sense, for example, to test some securities disclosure rules by randomly assigning individuals to different disclosure regimes. Instead, the policymaker would probably randomly assign firms to different disclosure regimes and observe how the different disclosure regimes affect firm outcomes. Alternatively, different jurisdictions might be assigned to different policies, with the same policy applying to each individual within a jurisdiction. If we wanted to examine the effect of different speed limits, for example, it would theoretically be possible to randomly assign every driver in the jurisdiction to a different speed limit and observe the outcome. But instead of giving each individual a different speed limit, policymakers could give different municipalities, counties, or states each a different

---

[187] This type of experiment is known as an "active-control trial." *See, e.g.*, Sharona Hoffman, *The Use of Placebos in Clinical Trials: Responsible Research or Unethical Practice?*, 33 CONN. L. REV. 449, 459 (2001) (distinguishing these trials from true placebo trials).

[188] *See* RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE 160-62 (2008) (explaining the main features of Medicare Part D).

speed limit, with a given limit applying to all individuals within the corresponding jurisdiction.

So how should policymakers determine the appropriate level of randomization? We believe the appropriate level of randomization is the smallest scale that still leaves interactions between the treated and untreated groups at a minimum. We generally prefer more fine-grained units of randomization if we are theoretically confident that the policy treatment will not impact the untreated group. When a policy targets individual incentives and has no "externalities"—effects that extend beyond an individual—then the treatment should be randomly assigned at the individual level. For example, *if* (counterfactually) individual driving patterns did not affect others, then different speed limits should be randomly assigned to different individuals. Assigning speed limits to broader-level jurisdictions under these conditions gains no benefit and limits the power of an experiment because it is much more costly to add observations.[189] Thus, random assignment to individuals would be the best strategy when a policy targets individual outcomes and there are no spillovers to other (untreated) individuals. However, in this driving example, it is probable that randomized speed limits may affect the driving patterns of the untreated drivers. There might generally be more accidents for both treated and untreated drivers if they drive at different speeds on the same highway. Drivers in the control group might be induced to drive more aggressively if they witness subject-group drivers going faster. Because of the strong possibility of these types of spillovers between the treated and untreated groups, it would be more appropriate to randomize speed limits at the jurisdiction level.

Randomization at the firm level is often the appropriate unit of analysis when analyzing policies that are dominantly targeted toward affecting firm behavior. Accordingly, randomized tests of corporate and securities law should often be implemented by randomly treating individual firms. But analogous concerns about spillover effects on untreated firms apply here as well. If treated firms are required to comply with an inefficient rule, then we should expect that untreated firms that need not comply with the rule would have a competitive ad-

[189] When policy is randomized at the state level, for example, "serial correlation" in error terms makes standard errors wide and therefore complicates the finding of statistically significant policy impacts. For details, see Marianne Bertrand et al., *How Much Should We Trust Differences-in-Differences Estimates?*, 119 Q.J. ECON. 249, 249-61 (2004).

vantage.[190] In equilibrium, we would expect the untreated firms to change their behavior: faced with weaker competitors, the untreated firms might increase their price or change the quality of their product. We might even see the advantaged, untreated firms expand their market share and stock price because of "losing" the treatment lottery. At times, the treatment-induced shift in market share may be relevant to evaluating the legal treatment itself. When the outcome in question concerns dimensions of social welfare that are not fully felt by the firms and their customers, however, the impact of the treatment on the untreated firm's behavior may undermine analysts' ability to parse the true causal mechanism. The presence of intra-industry competitive spillovers will often militate toward randomizing at the industry, instead of the firm, level.

After choosing the experimental population, experimenters must choose the appropriate duration for the experiment. Longer experimental periods offer some obvious advantages. A longer period increases the chance that the involved parties will become aware of the experiment and thus reduces the ability of the parties to avoid experimental effects by delaying behavior until the experiment is completed. Both factors mean that longer periods are more likely to provide better estimates of the true effects of an experimental policy than are shorter periods. At the same time, however, long-term experiments exacerbate the inequalities that experimentation creates. In addition, experimental policies will often prove to be failures, and lengthening the term of the experiment raises the cost of these failures. In total, the experimental period should be the shortest period necessary to obtain reasonably representative estimates of the true effects of the experimental policy.

In some circumstances, the length of the experiment will be contingent on the interim results of the experiment itself. As in drug testing, if the interim results point to a clear conclusion, it may be appropriate to shut down the study earlier than expected.[191] Once it

---

[190] *See* Richard Craswell, *Passing on the Costs of Legal Rules: Efficiency and Distribution in Buyer-Seller Relationships*, 43 STAN. L. REV. 361, 372-85 (1991) (discussing the market impact of efficient and inefficient mandates); *see also* Christine Jolls, *Accommodation Mandates*, 53 STAN. L. REV. 223, 230-72 (2000) (considering the efficiency of mandates intended to accommodate the special needs of population subgroups).

[191] For example, the National Institutes of Health shut down a study of the impact of circumcision on HIV infection rates in Africa when it discovered that circumcision had a significant protective effect. *See* Donald D. McNeil, Jr., *Circumcision's Anti-AIDS Effect Found Greater than First Thought*, N.Y. TIMES, Feb. 23, 2007, at A3 (noting that "two clinical trials were stopped . . . because the results were so clear").

becomes clear that one treatment is preferred to another, it is immoral and inefficient to expose subjects capriciously to the inferior policy. On the other hand, in some circumstances it will be appropriate to extend the length of the experiment to gather more information. In multilevel randomized testing, for example, follow-up testing of untested permutations may be warranted. Still, in other contexts it may be appropriate to continue the testing, but to alter the probable assignments of the different policy treatments. Google AdWords provides a potential example of this form of "convexification" in the context of Internet advertisements. If a randomized experiment initially suggests that, for example, "Tastes Great" is a more successful beer ad than "Less Filling," the Google software could automatically start increasing the probability that people will see the more successful advertisement.[192] This method—called "outcome-adaptive randomization"—mitigates the inefficiency of additional testing and allows the researcher to continue to collect some information on the longer-term effects of the various policy treatments.[193]

B. *Institution-Specific Guidelines*

The precise workings and advantages of randomized experimentation may differ greatly depending on whether a legislature or an administrative agency designs an experiment. Administrative law doctrine should tolerate the launch of randomized experiments, and once randomization becomes more common, an executive order might insist that agencies systematically consider what policies should be randomized. Meanwhile, there is a danger that legislators will ignore even solid evidence produced by randomized experiments, and this concern produces an argument for *self-executing* randomized experiments—where policy outcomes hinge directly on experimental results in a way specified in statutes. Agencies, by contrast, are less likely simply to ignore experimental results.

---

[192] For more information on Google AdWords, see *supra* note 118.

[193] *See, e.g.*, Ying Kuen Cheng et al., *Continuous Bayesian Adaptive Randomization Based on Event Times with Covariates*, 25 STAT. MED. 55, 56 (2006) ("[O]utcome-adaptive randomization . . . uses the data from patients treated previously in the trial to unbalance the randomization probabilities in favour of the treatment . . . observed to have comparatively superior outcomes. [It] provides a compromise between ethical concerns and the scientific goal of obtaining unbiased treatment comparisons.").

1. Administrative Agencies: The Case for a Randomization Impact Statement

Sometimes, as in the Medicare experiment, an agency may conduct an experiment as the result of a legislative decree, but it is also possible that an agency itself could decide to randomize policies. The courts would presumably examine such a decision with the usual tools of judicial review of administrative decisions, ensuring, for example, that the action was procedurally proper,[194] consistent with the law,[195] and representative of a permissible policy judgment.[196]

These hurdles should be straightforward for an agency to clear. As long as an agency goes through the ordinary notice-and-comment process—including providing a detailed explanation of an experiment's purpose in the notice of proposed rulemaking,[197] as well as a "concise, general statement" of basis and purpose[198]—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 legal regime nor the control legal regime is inconsistent with the agency's governing statute, a decision to launch an experiment should present no problem under *Chevron* review. Perhaps the most significant obstacle would be hard-look review, in which a court would examine the agency's justification for creating the experiment.[199] But

---

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

[195] *See, e.g.*, Chevron U.S.A., Inc. v. Nat'l Res. Def. Council, Inc., 467 U.S. 837, 842-45 (1984) (setting forth the contemporary standard of judicial review for evaluating agency interpretations of congressional mandates).

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

[197] *Id.* § 553(b) (requiring publication in the Federal Register of "general notice of proposed rulemaking" in most cases). Agencies often seek to meet the general notice requirement by publishing the actual rules they are considering, though even this level of notice is sometimes inadequate. *See, e.g.*, Portland Cement Ass'n v. Ruckelshaus, 486 F.2d 375, 402 (D.C. Cir. 1973) (remanding a case noting that the "record reveals a lack of an adequate opportunity . . . to comment on the proposed standards, due to the absence of disclosure of the detailed findings and procedures of the tests").

[198] 5 U.S.C. § 553(c) (requiring an agency to provide opportunity for comment by interested persons and to issue a "concise general statement of basis and purpose"). "Concise" and "general" are sometimes interpreted to mean "detailed" and "specific." *See, e.g.*, Auto. Parts & Accessories Ass'n v. Boyd, 407 F.2d 330, 338 (D.C. Cir. 1968) (warning against an "overly literal" interpretation of these words).

[199] *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 decision-making." (footnote omitted)).

hard-look review is supposed to be deferential,[200] 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 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 the notice-and-comment procedure to establish an experiment,[201] 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,[202] perhaps in part because this compromise would allow an administrative agency to renew a program indefinitely.[203] An agency should, however, at least be allowed to focus solely on the reason for conducting the experiment, rather than responding to comments on the merits of the underlying policy issue. Because an experiment produces data on 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.

---

[200] *See, e.g.*, Motor Vehicle Mfrs. Ass'n v. State Farm Mut. Auto. 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 agency.").

[201] This loosening of the requirement might occur when Congress has explicitly instructed an agency to conduct an experiment. Even here, however, the agency is likely to notify the public of its intent to run an experiment. For example, it may solicit third parties to perform the experiment. *See, e.g.*, 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, 8269 (Feb. 22, 2002) (soliciting "applications for demonstration projects that use disease management, along with coverage of prescription drugs, to improve the quality of services furnished to specific beneficiaries and to manage expenditures under Parts A and B of the Medicare program").

[202] The Administrative Procedure Act includes a general exemption from the notice-and-comment process "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, e.g.*, Tenn. Gas Pipeline Co. v. FERC, 969 F.2d 1141, 1145 (D.C. Cir. 1992) ("'[T]he limited nature of the rule cannot in itself justify a failure to follow notice and comment procedures.'" (quoting Council of the S. Mountains, Inc. v. Donovan, 653 F.2d 573, 582 (D.C. Cir. 1981))).

[203] *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, 377-78 (1989) (suggesting that courts treat a rule's temporary nature as a factor in determining whether it falls under the "good cause" exception in § 553(b)(B)).

To enact an experimental policy permanently, an agency presumably would face hard-look review, but here too courts should be more deferential than usual. Critics of the notice-and-comment process complain that it has "ossified" the rulemaking process,[204] 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.[205] But an agency conducting an experiment is less likely to be following an ideological agenda than an agency drawing inferences based on existing data that might plausibly support a variety of conclusions. Moreover, courts conducting judicial review should recognize the unique value of evidence from randomized experimentation.[206] There remains a danger that an agency might make invalid inferences on the basis of an experiment. At least when 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 all of the problems of nonexperimental evidence on a case-by-case basis.

If individual agency initiatives begin to test and choose policies with experimental means, randomization could gradually become a more entrenched part of the policymaking process. Perhaps, considering randomization might become almost as routine and formalized as a cost-benefit analysis.[207] For example, the executive branch could provide a standard procedure for agencies to consider randomization and to produce a randomization impact statement (RIS) when enacting a new rule—regardless of whether the agency decided to use a randomized approach.

An RIS might include the following elements:

---

[204] Thomas O. McGarity, *Some Thoughts on "Deossifying" the Rulemaking Process*, 41 DUKE L.J. 1385, 1438 (1992); *see also* Richard J. Pierce, Jr., *Seven Ways to Deossify Agency Rulemaking*, 47 ADMIN. L. REV. 59, 61 (1995) (observing that even though agencies should try to keep up with scientific and technological advances, "agencies rarely amend rules because the amendment process is as daunting as the process of promulgating a rule").

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

[206] *Cf.* Dorf & Sabel, *supra* note 2, at 397 (arguing that hard-look review should reward experimental agency approaches with greater deference (though not focusing specifically on random experimentation)).

[207] *See* Exec. Order No. 12,866, 3 C.F.R. 638, 639 (1994), *reprinted as amended in* 5 U.S.C. 601 (2006) (permitting regulation only where benefits exceed costs).

1. *The impetus for conducting a policy experiment.* It will be particularly important to delineate the particular predicted outcomes or consequences that motivate the proposed change. If no experiment was conducted, an explanation of the experiment's absence should be provided. Valid explanations for the absence of an experiment would include a de minimis exception, overwhelming evidence about the policy's desirability, an urgent need for a new policy, or the impossibility of conducting a truly informative experiment. In some circumstances, it will prove difficult to measure quantitatively the information about the impacts of interest or to do so in a timely fashion. At other times, it will prove impossible to reach a consensus about how to weigh the importance of various impacts. For example, we imagine that a randomized experiment looking at the impact of a spousal-notification requirement for abortion might do little (even if such a test were constitutionally permissible)[208] to resolve the legislative debate because legislators and their constituent groups may have incommensurable preferences.[209]

2. *A detailed description of the experiment.* The description should discuss the unit of randomization, the scope and length of the experiment, and the anticipated possible effects of the experimental policy on different outcome measures.

3. *A summary of the results of the experiment.* The summary should reflect not only the agency's examination of the data the experiments

---

[208] *See* Planned Parenthood of Se. Pa. v. Casey, 505 U.S. 833, 898 (1992) (striking down a Pennsylvania spousal-notification law). But a state might experiment with offering couples at the time of marriage the option of contracting for spousal notification. *See* Andrew Blair-Stanek, Comment, *Default and Choices in the Marriage Contract: How to Increase Autonomy, Encourage Discussion, and Circumvent Constitutional Constraints,* 24 TOURO L. REV. 31, 49-51 (2008).

[209] Then again, some moderate legislators might be swayed by compelling evidence about the impact of notification law on (1) a woman's propensity to abort; (2) the propensity of those who would have otherwise been aborted fetuses to commit crime; and (3) the probable psychological well-being of the spouses. *Cf.* Donohue & Levitt, *supra* note 61, at 380 (2001) ("[W]e consider a novel explanation for the sudden crime drop of the 1990s: the decision to legalize abortion over a quarter century ago."); Cass R. Sunstein & Adrian Vermeule, *Is Capital Punishment Morally Required? Acts, Omissions, and Life-Life Tradeoffs*, 58 STAN. L. REV. 703, 705 (2005) (arguing that statistical evidence on the deterrent effect of the death penalty, if it exists, may make the death penalty a moral imperative); Cass R. Sunstein & Justin Wolfers, Op-Ed., *A Death Penalty Puzzle: The Murky Evidence for and Against Deterrence*, WASH. POST, June 30, 2008, at A11 (calling for public debate on the death penalty to avoid distortions based on "misunderstanding . . . what the evidence actually shows").

generate but also the analysis of other researchers. If there are differences of opinion regarding the outcomes of the experiment, the RIS should discuss reasons for the differences and explain why the agency prefers one conclusion about the causal effects of a policy over another.

4. *An explanation of why the results weigh in favor of adopting a new policy.* The results of the experiment are simply data. The results provide information that informs policymaking, but they cannot specify how policymakers should prefer certain outcomes over others. Consequently, the RIS should explain why the causal impacts of the policy are desirable in light of the stated goals of the agency.

An important question would be the role of the courts in reviewing RISs. Once again, no doctrinal innovation is necessary here, as the courts could assess RISs with the usual tools of hard-look review, ensuring that agencies have carefully addressed counterarguments both to decisions about whether to engage in randomization and to decisions after experimentation occurs. Creation of the RIS as an integral part of the administrative process would ensure that consideration of randomization by both agencies and courts would become a standard part of the policy process, rather than an occasional innovation pushed largely by academic researchers.

The RIS could also provide an opportunity for the Office of Management and Budget (OMB), or some other specialized agency, to generate expertise in administering experiments on which all agencies could rely. The OMB, or another specialized agency, might even be given the general task of conducting all policy experiments and interpreting results. Running policy experiments requires specific skills, such as knowing what types of outcome information are readily obtainable and limiting dropout rates in the subject and control populations. Many of these skills apply regardless of the subject of the policy experiment, and there is likely to be considerable learning by doing. Just as pharmaceutical companies hire clinical trial companies to run drug trials, so too should policymaking bodies use experimental-trial specialists.[210]

[210] One prominent clinical-trial company has run over "3,200 trials in some 90 countries" since 2000. QUINTILES TRANSNATIONAL CORP., THE VALUE OF CONTRACT RESEARCH ORGANIZATIONS 2, http://www.quintiles.com/elements/media/white-papers/value-cros.pdf (last visited Jan. 15, 2011).

2. Legislatures: The Case for Self-Execution

Given the problems identified in Part III, random-experimentation data will rarely give unambiguous answers to multidimensional policy questions. The results of random experimentation will become 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.[211] But experiments could have greater impact if legislatures were to make policy conditional on experimental results—that is, if experiments were self-executing. 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 in an administrative agency to make policy changes based on the experiment. 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 authorization, and it therefore cannot avoid these obstacles altogether. But if through gradual steps randomized self-executing experiments become sufficiently familiar that they no longer seem strange, 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 a law's staunchest opponents might nonetheless be willing to support the experiment if they believe that the experiment will prove the law 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, a bidirectional self-executing experiment, which would move the law automatically in the direction of the experiment if it proves the law to be successful, and in the opposite direction if it proves the law to be a failure. If ideological opponents genuinely disagree about the effects of potential policies, such experiments can seem beneficial ex ante,

---

[211] *See supra* notes 43-45 and accompanying text (discussing an example of policymakers misunderstanding certain data).

increasing the gains from political trade. Such experiments also channel ideological disagreement, increasing the possibility of legal change, rather than obstruction. Finally, while these experiments seem unlikely in our current legal system, increased comfort with experimentation, randomization, and self-execution might someday create a public that perceives opposition to experiments as an indication that policymakers lack confidence in the empirical validity of their proposals.

The principal challenge of self-execution is determination of what counts as success. The metrics upon which policymakers agree may well be simple proxies—far less sophisticated than what statisticians would rely on ex post. For example, success might depend on a comparison of a single variable, or perhaps two or three variables, between the treatment and control group. In theory, policymakers might agree in advance on regression designs and a formula aggregating regression coefficients or other results to create nuanced self-executing experiments. But it is difficult to conceive in advance all of the regression tests that would be necessary to verify robustness. Any formula is likely to be somewhat arbitrary and difficult to understand. 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, relying on simple proxies for determining the success of randomized self-executing experiments may be politically more feasible.

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

Ultimately, 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. In general, simple designs will be preferable to more complex and sophisticated ones when there is a danger either that the relevant officials will be unable to agree on the policy significance of a randomized experiment, or that, even if some authoritative decisionmaker could reach a resolution, such a decisionmaker might be biased in favor of a particular policy preference.

C. *Applications*

To demonstrate the generality of such experiments, in this Section we develop policy-experiment applications to different fields of law and policy, focusing on some fields that have not been considered as candidates for randomized experimentation in the past. We begin with a detailed examination of an actual randomized securities law experiment and propose extending the same approach to test the Sarbanes-Oxley Act. We then consider the possibility of a randomized test in the area of taxation.

1. Securities Law

Securities law is ideally situated for randomized policy experiments. Much of securities law applies at a national level. As a result, there is little interstate variation in securities law that scholars can apply to test different approaches to securities regulation.[212] Moreover,

---

[212] The lack of interstate variation explains the intense empirical interest in the relatively infrequent change in securities law at the national level. For example, the 1930s, when modern securities law was first introduced, continues to constitute an active area of research, as do the 1960s, when an expansion of the securities law regime to over-the-counter (OTC) stocks occurred. *See, e.g.*, Michael Greenstone et al., *Mandated Disclosure, Stock Returns, and the 1964 Securities Acts Amendments*, 121 Q.J. ECON. 399, 400 (2006) (noting that "the 1964 [Securities Acts] Amendments provide a compelling setting for evaluating the consequences of mandatory disclosure regulations"); Paul G. Mahoney, *The Political Economy of the Securities Act of 1933*, 30 J. LEGAL STUD. 1, 1 (2001) (describing the 1933 Securities Act as "a severe testing ground for rent-seeking theories of economic regulation"); Allen Ferrell, *Mandated Disclosure and Stock Returns: Evidence from the Over-the-Counter Market* 6-33 (Harvard John M. Olin Center for Law, Economics, and Business, Discussion Paper No. 453, 2003), *available at* http://ssrn.com/abstract=500123 (using the 1964 Amendments to study the effect of imposing mandated disclosure on the informational efficiency of the OTC market).

many topics in securities regulation, such as the desirability of short selling or the appropriate degree of required disclosure, are the subject of long-standing, but still hotly contested, debates.[213] Securities regulation is characterized by intense theoretical debates informed by scant empirical evidence. Systematic randomized policy experiments offer the prospect of providing important new data to many of these traditional, theoretical debates.

a. *A Short-Sale Experiment*

Policymakers recently have begun to grasp the potential of randomized policy experiments for securities. In 2004, the Securities and Exchange Commission (SEC) issued Rule 202T of Regulation SHO, devising an experiment to test some restrictions on short sales.[214] Scholars have debated the effect of these restrictions. Finance theory predicts that short-sale restrictions should reduce the volume of short selling, an effect that, in turn, should reduce the liquidity of a stock and potentially lead to less accurate pricing.[215] Others argue that the restrictions help to prevent coordinated short-sellers—seeking to force the price of a stock down simply to purchase it at a low price—from manipulating stock prices.[216]

Rule 202T allowed the SEC to implement a "pilot program to examine the efficacy" of the short-sale restrictions.[217] The pilot program exempted one-third of the stocks in the Russell 3000, an equity index,

---

[213] *See, e.g.*, Stephen E. Christophe et al., *Short-Selling Prior to Earnings Announcements*, 59 J. FIN. 1845, 1847 (2004) ("[R]egulators should consider requiring markets to make more extensive and timely disclosures of short-selling activity."); Ian Ramsay, *Short Selling: Further Issues*, 21 SEC. REG. L.J. 214, 218 (1993) (arguing that a short-selling disclosure rule could "impede market efficiency" and may not operate quickly enough to allow information to "disseminate into the market").

[214] Short Sales, 69 Fed. Reg. 48,008, 48,012 (Aug. 6, 2004); *see also* Regulation of Short Sales, 17 C.F.R. § 242.200–.204 (2010); OFFICE OF ECON. ANALYSIS, U.S. SEC. & EXCH. COMM'N, ECONOMIC ANALYSIS OF THE SHORT SALE PRICE RESTRICTIONS UNDER THE REGULATION SHO PILOT 3-4 (2007), *available at* http://www.sec.gov/news/studies/2007/regshopilot020607.pdf (describing the restrictions in place before Rule 202T was enacted and the purpose of the rule itself).

[215] *See* OFFICE OF ECON. ANALYSIS, *supra* note 214, at 6-8 ("Finance theory predicts that under certain conditions, constraints on short selling may cause securities to be misvalued by the market, particularly when investors have highly divergent opinions about the stock.").

[216] *See* Emilios Avgouleas, *A New Framework for the Global Regulation of Short Sales: Why Prohibition Is Inefficient and Disclosure Insufficient*, 115 STAN. J.L. BUS. & FIN. 376, 380-83 (2010) (describing and countering the market-abuse argument offered by the government for short-selling restrictions).

[217] *Id.* at 4.

from the short-sale restrictions.[218] The exempted stocks were chosen "by sorting the 2004 Russell 3000 first by listing market [e.g., NYSE, NASDAQ] and then by average daily dollar volume from June 2003 through May 2004, and then within each listing market selecting every third company starting with the second."[219] This is an example of stratified sampling.[220] So long as it is effectively random which of the three companies with similar daily trading volumes happens to get exempted from the restrictions, the selection mechanism is equivalent to a stratified randomized experiment. Note that the SEC's experimental design in this case did not seek volunteer companies for different regimes. Instead, the SEC simply chose some companies that would be exempted from the current short-sale restrictions.

The exempted stocks and the other stocks in the Russell 3000 operated under different trading regimes from May 2005 to August 2007, providing a significant period for observing the effects of the short-sale restrictions relative to eliminating the restrictions.[221] The Office of Economic Analysis of the SEC produced a comprehensive report on the pilot program, including many of the components that we recommend for the RIS. The report first reviews the theoretical and empirical literature on short-sale restrictions.[222] This literature tends to view the existing policy of short-sale restrictions as ineffi-

---

[218] *Id.*

[219] *Id.* n.6.

[220] Stratified sampling occurs because

> [i]n any randomized trial it is desirable that the comparison groups should be as similar as possible as regards participant characteristics that might influence the response to the intervention. Stratified randomization is used to ensure that equal numbers of participants with a characteristic thought to affect prognosis or response to the intervention will be allocated to each comparison group. . . . Stratified randomization is performed either by performing separate randomization (often using random permuted blocks) for each strata, or by using minimization.

Evidence-Based Medicine, SA HEALTHINFO, http://www.sahealthinfo.org/evidence/s.htm (last visited Jan. 15, 2011). If trading volume influences the effect of short-sale restrictions, then the pilot design insured that the exempt group of stocks and the control group were similar by performing separate selections for each group of three stocks with similar daily trading volumes.

[221] *See* OFFICE OF ECON. ANALYSIS, *supra* note 214, at 4 ("The Pilot went into effect on May 2, 2005, and was scheduled to end on April 28, 2005, but [was] extended to August 6, 2007, to allow the Commission to consider potential rulemaking after evaluating the results of the Pilot." (footnote omitted)).

[222] *Id.* at 16-22.

cient.[223] The report explains that the pilot program was enacted "to obtain empirical data to help assess whether short sale regulation should be removed, in part or in whole, for actively-traded securities, or if retained, should be applied to additional securities."[224] The report also provides detailed descriptions of the possible effects of short-sale restrictions on a wide variety of outcomes, such as short-selling volume, the amount of "synthetic" short sales in the option markets or via trading platforms, liquidity, pricing levels, and pricing volatility.[225]

The report then explains how the experiment was conducted, with a discussion and justification of the stratified sampling method used in the experiment.[226] In addition, the report explains the methodological tools applied to examine the impact of the short-sale restrictions on various outcomes.[227] Finally, the report examines the impact of the short-sale restrictions on the outcomes of interest—including short-selling volumes, bid-ask spreads, and use of short-sale substitutes, such as put options.[228] The report examines each outcome variable of interest and finds that eliminating short-sale restrictions affects some outcome variables (such as short-selling volumes, which are approximately eight percent less with the restrictions than without) but has no effect on others (there are no differences in bid-ask spreads with or without the restrictions).[229] The report also describes other studies of the pilot program's experimental elimination of short-sale requirements and discusses differences in estimated effects between the SEC's study and the other academic studies.[230] The report concludes,

> In summary, having examined the impact of the Regulation SHO Pilot on a wide array of market characteristics, we conclude that price restrictions constitute an economically relevant constraint on short selling. Our evidence suggests that removing price restrictions for the pilot stocks has had an effect on the mechanics of short selling, order routing

---

[223] *See id.* at 22 ("Overall, this evidence seems to indicate that tick tests can lead to narrower bid ask spreads, but impedes price discovery, while the bid tests should not have any discernible effect on market quality.").

[224] *Id.* at 4 (quoting Order Suspending the Operation of Short Sale Price Provisions for Designated Securities and Time Periods, Exchange Act Release No. 50,104, 69 Fed. Reg. 48,032, 48,032 (Aug. 6, 2004)).

[225] *Id.* at 6-9.

[226] *Id.* at 22-27.

[227] *Id.* at 28-34.

[228] *Id.* at 34-51.

[229] *Id.* at 51-57; *see also id.* at 62 tbl.3, 65 tbl.6.

[230] *Id.* app. A.

> decisions, displayed depth, and intraday volatility, but on balance has not had a deleterious impact on market quality or liquidity.[231]

The report does not go beyond these conclusions to suggest policy changes in response to the experiment, although any subsequent attempt to change short-sale restrictions is likely to discuss the pilot program in detail.

In total, the nearly randomized elimination of short-sale restrictions for one-third of the firms in the Russell 3000 highlights the value of experiments for policymaking. The experiment demonstrated that short-sale restrictions have some effects in the predicted direction, such as a reduction in short-selling volume, but that it is unlikely that elimination of the restrictions would have a dramatic effect on market efficiency. Such sober conclusions suggest that experiments do not always lead to dramatic outcomes. On the one hand, advocates of repeal can argue that the short-sale restrictions reduce freedom without producing any demonstrable improvement in market efficiency. Increasing individual freedom without hurting others presents a strong case for repeal. On the other hand, advocates of the status quo can argue that some of the benefits of the restriction—particularly the possibility of stabilizing the market during a price meltdown—were not amenable to easy testing. Moreover, the costs of the restrictions are small. The restriction has no systematic effect on bid-ask spreads. With relatively low costs and untested benefits, proponents of the short-sale restriction can argue that the case for repeal has not been made. At a minimum, the existence of the randomized test results makes some of the more strident arguments for and against repeal of the short-selling restrictions less plausible.

The quality of the experimental short-sale restriction elimination and its accompanying report raises an obvious question. Given how valuable the experiment appears to be and how efficiently it was conducted, why does the SEC not apply its experimental expertise systematically to other debates in securities regulation? The next subsection proposes such an experiment in one area—the Sarbanes-Oxley Act—but experiments can apply to any controversial issue.

b. *Experimental Sarbanes-Oxley Repeal*

In the wake of the Enron and WorldCom accounting scandals in 2002, Congress passed the Sarbanes-Oxley Act.[232] Sarbanes-Oxley in-

---

[231] *Id.* at 56.

cluded many provisions to improve the quality of financial reporting and corporate governance. Some of Sarbanes-Oxley's prominent provisions include mandatory CEO and CFO certification of financial results[233] and new "internal-controls" requirements.[234]

Sarbanes-Oxley has proven controversial. Many corporations and academics dispute Sarbanes-Oxley's efficacy in preventing fraud, while bemoaning its expense.[235] Others argue that Sarbanes-Oxley performs a critical role in improving confidence in financial markets.[236] This debate has spawned an extensive empirical literature evaluating Sarbanes-Oxley's impact on corporate value, cross-listing in the U.S. markets, and going-private decisions.[237] Many empirical papers find that Sarbanes-Oxley appears to destroy value or reduce cross-listings, but others dispute these findings.[238]

The ambiguity about Sarbanes-Oxley's desirability is reflected in calls for its elimination.[239] To this point, however, Sarbanes-Oxley's proponents have managed to prevent its alteration. Sarbanes-Oxley, then, offers an almost ideal context for a randomized repeal of securities legislation. Sarbanes-Oxley's provisions may well destroy value,

---

[232] Sarbanes-Oxley Act of 2002, Pub. L. No. 107-204, 116 Stat. 745 (codified as amended in scattered sections of 11, 15, 18, 28 and 29 U.S.C.).

[233] 15 U.S.C. § 7241 (2006).

[234] 15 U.S.C. § 7262. The internal-controls requirements obligated companies to set up elaborate mechanisms for detecting malfeasance within the company or disclose the absence of such controls.

[235] *See, e.g.*, Peter C. Kostant, *From Lapdog to Watchdog: Sarbanes-Oxley Section 307 and a New Role for Corporate Lawyers*, 52 N.Y.L. SCH. L. REV. 535, 536 n.2 (2008) (collecting sources criticizing the Act).

[236] *See id.* at 536-37 (defending, in particular, section 307).

[237] *See, e.g.*, Ellen Engel et al., *The Sarbanes-Oxley Act and Firms' Going-Private Decisions*, 44 J. ACCT. & ECON. 116, 143 (2007) ("[S]maller firms with high inside ownership experience higher going-private announcement returns in the post-SOX period compared to the pre-SOX period."); Peter Iliev, *The Effect of SOX Section 404: Costs, Earnings Quality, and Stock Prices*, 65 J. FIN. 1163, 1193 (2010) (finding that for small firms the cost of Sarbanes-Oxley section 404 outweighs the benefits); Roberta Romano, *The Sarbanes-Oxley Act and the Making of Quack Corporate Governance*, 114 YALE L.J. 1521, 1529 (2005) (arguing for the removal of mandatory corporate governance provisions from Sarbanes-Oxley); Ivy Xiying Zhang, *Economic Consequences of the Sarbanes-Oxley Act of 2002*, 44 J. ACCT. & ECON. 74, 110 (2007) ("[T]he cumulative abnormal returns of U.S. firms and foreign firms complying with SOX around key SOX events are negative and statistically significant.").

[238] *See generally* Christian Leuz, *Was the Sarbanes-Oxley Act of 2002 Really this Costly? A Discussion of Evidence from Event Returns and Going-Private Decisions*, 44 J. ACCT. & ECON. 146 (2007) (discussing and contributing to this academic debate).

[239] *See* Romano, *supra* note 237, at 1529 (arguing that "the corporate governance provisions of SOX should be stripped of their mandatory force and rendered optional for registrants").

but the existing empirical evidence is difficult to interpret because of confounding factors that plague the studies. For example, foreign company cross-listings in U.S. markets may have declined because of Sarbanes-Oxley's onerous requirements, or they may have declined due to the development of sophisticated foreign exchanges, decreasing the value of U.S. markets as a source of capital. An experimental repeal of Sarbanes-Oxley for some companies is likely to provide convincing empirical evidence that resolves which of these factors is more important. Moreover, because Sarbanes-Oxley is so unpopular with corporations, instituting an experimental repeal should prove popular, while avoiding the political battle that attempting to repeal Sarbanes-Oxley permanently for all companies would cause.

Randomized experimental repeal of Sarbanes-Oxley should take place as follows. First, the most controversial provisions of Sarbanes-Oxley should be identified. These are likely to include the internal-control provisions and the CEO and CFO certification provisions. These provisions should then be randomly repealed for some corporations. The randomization should be stratified to ensure that different types of companies are appropriately represented in both the treatment group (with the Sarbanes-Oxley restrictions repealed) and the control group (with Sarbanes-Oxley continuing as presently). For example, foreign companies cross-listed in U.S. markets should be well represented in both the sample and the control group to help evaluate Sarbanes-Oxley's effect on delisting from U.S. markets.

The experimental repeal period should be relatively long. Many of Sarbanes-Oxley's effects will be felt only gradually. Corporate fraud, for example, does not occur overnight. In addition, once a plan for internal controls has been disbanded, it requires significant time and expense to restart it. In response, companies subject to experimental repeal will not scrap or revise their costly internal-control mechanisms unless they can be confident that they will not have to reinstate the mechanisms shortly thereafter. As a result, a short-term experimental Sarbanes-Oxley repeal will not provide a good test of Sarbanes-Oxley's true effects.[240] Instead, the experimental repeal should be applied for an extended period—up to several years.[241]

---

[240] Because market values incorporate expectations of future profits, market values respond very quickly to new policies. The magnitude of the response to a new policy, however, will depend upon the policy's duration, as well as the policy's expected impact. A short-term experimental repeal of Sarbanes-Oxley may therefore have a small (and potentially indistinguishable) effect on corporate value because the experiment will not take place over a long enough period to have an important effect on long-term

Just as in the short-sale experiment, the unit of observation for an experimental Sarbanes-Oxley repeal should be the publicly traded company. Sarbanes-Oxley's requirements apply to publicly traded corporations, making the choice of unit of observation relatively straightforward. If the repeal of Sarbanes-Oxley is likely to produce substantial competitive advantages for untreated firms (that is, those still subject to Sarbanes-Oxley requirements), then the unit of randomization may need to be raised to the industry level.[242] Even the possibility of being put at a competitive disadvantage might make industry randomization politically more palatable.

The randomization should occur on each controversial issue within Sarbanes-Oxley, rather than on Sarbanes-Oxley as a whole. Thus, some companies would have the internal-control provisions eliminated, but other provisions of Sarbanes-Oxley would remain intact. Others would have only the CEO and CFO certification provisions eliminated. Still others would have both these provisions eliminated but the rest of Sarbanes-Oxley intact, and so on. Randomizing different permutations of the controversial provisions in Sarbanes-Oxley allows for the identification of specific provisions that are effective or ineffective, rather than attempting to judge the law as a whole. In addition, observing the effects of different permutations allows policymakers to see if there are any interaction effects between the two provisions.[243]

---

profitability. Moreover, market responses, even if correct in expectation, may prove wrong in reality. A longer-term experiment allows the researchers to determine actual effects, rather than simply anticipated effects.

[241] While several years may sound like a long period, the status quo, with a controversial law applied indefinitely, is in many ways just as speculative an experiment, but one that does not produce information that would yield policy conclusions.

[242] For example, suppose that investors benefit from the improvement in information quality mandated by Sarbanes-Oxley, but that investors can apply this information from companies subject to Sarbanes-Oxley to companies not subject to Sarbanes-Oxley. In this case, the non-Sarbanes-Oxley companies may do better than the Sarbanes-Oxley companies because they get the benefit of the improved information without incurring its expense. This difference in outcomes, however, does not accurately estimate the effects of a full Sarbanes-Oxley repeal. If no companies followed Sarbanes-Oxley, then there would be no informational spillovers, and all companies might be worse off. An experiment that is partially randomized at the industry level and partially randomized at the firm level could parse out the extent to which there were intra-industry spillovers of this kind.

[243] An interaction effect occurs when the effect of one variable depends on the value of another variable. For example, CEO-certification provisions taken alone might not impact corporate value. Similarly, internal-control requirements taken alone may also have no impact on value. When the two provisions are implemented together, however, they may have mutually reinforcing effects; the combination of the two provisions would then have an impact on value.

Because many companies find Sarbanes-Oxley compliance costly and are likely to volunteer, policymakers could ask for companies to volunteer to participate in a Sarbanes-Oxley-repeal experiment and then could assign some of these companies to a Sarbanes-Oxley-repeal treatment group and others to a control group with Sarbanes-Oxley remaining in place.[244] The experiment with volunteer companies would provide a good estimate of the treatment effect of allowing companies to opt out of Sarbanes-Oxley because companies that volunteer to take part in an experimental repeal are likely to be similar to companies that would opt out of Sarbanes-Oxley were that an option. Examining an experiment with volunteers would provide a poor estimate of the effect of a full repeal of Sarbanes-Oxley, however, because the impact of Sarbanes-Oxley on companies that volunteer to have it eliminated is likely to be different from the impact of Sarbanes-Oxley on the average company.[245]

To estimate the impact of a full Sarbanes-Oxley repeal on the average company, the repeal could be randomly, but mandatorily, assigned to some companies but not to others. This method would incur the cost of forcing some companies to experience Sarbanes-Oxley repeal unwillingly, but it would avoid the problem of estimating the impact of Sarbanes-Oxley exclusively for companies that volunteer to have it repealed. A randomized mandatory repeal of Sarbanes-Oxley for some companies but not for others is no different than the randomly assigned repeal of short-sale restrictions undertaken in the Regulation SHO pilot. An intermediate strategy would be to randomize all companies except those that decide to opt out of the experiment, ensuring that failure to act is not interpreted as unwillingness to participate in the experiment.

---

[244] Repealing Sarbanes-Oxley for all companies that volunteer for the Sarbanes-Oxley-repeal experiment and estimating the impact of Sarbanes-Oxley by comparing these companies with companies that did not volunteer for the experiment (for whom Sarbanes-Oxley remained in place) fails to provide accurate estimates of the impact of Sarbanes-Oxley. Companies that volunteer for Sarbanes-Oxley repeal may be different in unobservable ways from companies that do not volunteer. Any differences in outcomes for the two groups may therefore be attributable to these unobserved differences rather than to the repeal of Sarbanes-Oxley. As a result, some companies that volunteer for Sarbanes-Oxley repeal should be randomly assigned to a control group that must remain compliant with Sarbanes-Oxley. These companies will be similar to the companies that volunteered for a Sarbanes-Oxley repeal and were randomly assigned to the group that no longer was required to remain compliant, making estimates of the effect of a Sarbanes-Oxley repeal more accurate.

[245] *See supra* note 89 and accompanying text (providing an analytical background for experiments conducted on self-selecting groups).

There are many potential outcomes of interest for a Sarbanes-Oxley randomized experiment. Sarbanes-Oxley aimed to restore investor confidence in the financial markets and financial reporting. Therefore, one obvious outcome variable is investor confidence in the quality of corporate reporting. A related measure would include the incidence of fraud in Sarbanes-Oxley companies relative to non-Sarbanes-Oxley companies. To financial economists, however, confidence and prevention of fraud are not aims but rather means to an end.[246] Investor confidence should reduce the cost of equity and debt financing, thereby enabling more investment in positive net-present-value activities. Moreover, measures of investor confidence or fraud prevention fail to account for the cost of Sarbanes-Oxley compliance. Therefore, other measures that account for both the costs and benefits of Sarbanes-Oxley should be examined.

One important alternative measure of Sarbanes-Oxley's efficacy is stock market value. Stock market value goes up if investors perceive that Sarbanes-Oxley reduces the cost of capital without costing anything itself, but it goes down if Sarbanes-Oxley raises costs without benefits. The stock market response to the announcement of the randomization status of each company will therefore provide a good estimate of the market's impression of Sarbanes-Oxley's net effects. Because of the randomized nature of a Sarbanes-Oxley experiment and the large number of companies that would participate, a long-term study of the impact of Sarbanes-Oxley on market value is possible. Such a study would provide evidence not just of the market's impressions of Sarbanes-Oxley, but also of the market's verdict after observing Sarbanes-Oxley's impacts. If, after a number of years, Sarbanes-Oxley companies have outperformed non-Sarbanes-Oxley companies, then this result would constitute solid evidence that Sarbanes-Oxley enhances corporate value.

If a Sarbanes-Oxley experiment is to be self-executing, simple comparisons of stock market values for treated and control corporations may be the best basis for determining whether Congress should retain particular features of Sarbanes-Oxley. The case for self-execution is particularly strong if it appears likely that Congress otherwise might ignore the experiment, with partisans sticking to their orig-

---

[246] *See* Irwin H. Steinhorn & William M. Lewis, *Corporate Compliance Under the Regulations Implementing Sarbanes-Oxley*, 60 CONSUMER FIN. L.Q. REP. 30, 30 (2006) (noting that "'[t]he strength of the U.S. financial markets depends on investor confidence'" (quoting Disclosure Required by Sections 404, 406, and 407 of the Sarbanes-Oxley Act of 2002, Securities Act Release No. 8177, 68 Fed. Reg. 5110, 5110 (Jan. 31, 2003))).

inal positions regardless of the experiment's results. Even a perfect experiment can not resolve all questions about Sarbanes-Oxley. For example, partisans might argue reasonably that the result could have been different if the experiment had lasted longer. It might seem that an experiment's imperfection furnishes an argument *against* self-execution, on the ground that policy changes should depend on ex post expert analysis. Arguably, though, imperfection furnishes an argument *for* self-execution, if a proxy result is still meaningful and legislators seem likely to have "sticky priors."[247] An imperfect proxy may be more likely to produce beneficial legislative change than might careful analysis if legislators seem unlikely to be swayed by such analysis. In any event, self-execution would merely change the policy baseline; Congress could still act based on a nuanced interpretation of the experiment.

In addition to running tests on Sarbanes-Oxley, the SEC could run analogous experiments that investigate other contentious issues in securities law, such as whether mandatory disclosure or insider trading prohibitions enhance corporate value, or merely add costs. Such experiments should follow the format we suggest here for Sarbanes-Oxley, which, in turn, is very similar to the experimental short-sale restriction study already run by the SEC.

2. Tax Law

Few topics in public policy are as hotly debated as the impact of different tax rates on incentives to work. Some economists argue that small changes in marginal tax rates can have large effects on work hours and entrepreneurship. As a result, they claim that lowering marginal tax rates does not reduce government revenues as much as one might predict.[248] Others argue that hours and entrepreneurship are not particularly sensitive to relatively small changes in marginal tax rates, meaning that government revenues will fall nearly propor-

---

[247] For a discussion of "sticky priors," see generally Lisa E. Bolton & Americus Reed II, *Sticky Priors: The Perseverance of Identity Effects on Judgment*, 41 J. MARKETING RES. 397 (2004).

[248] If a change in tax rates has no impact on behavior, then the revenue loss can be estimated by the decrease in the tax rate. Most economists, however, think that a change in the tax rate has some effect on the supply of labor and entrepreneurship. Some economists even claim that lowering tax rates can increase revenue, but this claim is discredited. *See* N. Gregory Mankiw, *The Optimal Collection of Seigniorage: Theory and Evidence*, 20 J. MONETARY ECON. 327, 332 (2004) (suggesting that an increase in tax revenue is normally associated with higher taxes).

tionately to the amount of a tax decrease.[249] These arguments are rehashed whenever the government considers raising or lowering taxes—in other words, almost annually.[250]

Because tax rates change frequently, there is ample variation with which to study how the change in tax rates impacts labor supply and entrepreneurship.[251] Unfortunately, these changes in rates are often correlated with many other changes, making it extremely difficult to draw firm conclusions about the response of labor supply to tax rates.[252] For example, tax rates are often altered in response to changes in economic conditions.[253] If economic behavior changes after rates change, the behavior changes may be attributable to the change in rates, or they may be attributable to the altered economic conditions that motivated the change in rates in the first place. Such confounding factors help explain the lack of consensus about the true impact of taxes on labor supply incentives.[254]

Randomized experimental manipulation of tax rates will not suffer from this complication. If tax rates are randomized at the individual level, then individuals facing similar economic conditions will be subject to different tax rates. If these individuals behave differently, then

---

[249] *See* Paul Krugman, *The Laffer Test (Somewhat Wonkish)*, THE CONSCIENCE OF A LIBERAL, N.Y.TIMES.COM BLOG (Aug. 10, 2010, 10:37 AM), http://krugman.blogs.nytimes.com/2010/08/10/the-laffer-test-somewhat-wonkish (explaining that even very high marginal tax rates are not likely to reduce work much, though they do increase incentives to evade taxes).

[250] *See, e.g.*, Glenn Kessler, *Now President Faces Tax Cut Test: Loss of Revenue Means Bush Needs to Slow Spending*, WASH. POST, Feb. 11, 2001, at A5 (suggesting that a tax cut must be accompanied by a reduction in spending); David E. Rosenbaum, *Name That Tune About Tax Cuts*, N.Y. TIMES, May 18, 2003, at BU4 (reporting Congress's Joint Committee on Taxation's finding that the "revenue feedback" from proposed tax cuts would be approximately three to twenty-three percent of the cuts over ten years).

[251] *See, e.g.*, Daniel J. Mitchell, *Lowering Marginal Tax Rates: The Key to Pro-Growth Tax Relief*, BACKGROUNDER (Heritage Foundation, Washington, D.C.), May 22, 2001, at 1, 5-7 (using the Kennedy tax cuts and the Reagan tax cuts to argue that lower taxes stimulate economic growth, leading to higher revenue).

[252] *See* Basil Dalamagas, *The Effects of Tax Rate Changes on Output and Government Deficits*, 10 APPLIED ECON. LETTERS 97, 101 (2003) (concluding that the effects of lowering the tax rate vary based on "the kind of effective tax rates which fiscal authorities choose to decrease and the ratio of factor income tax rate to consumption tax rate").

[253] *See, e.g.*, David M. Herszenhorn, *Bush and House in Accord for $150 Billion Stimulus*, N.Y. TIMES, Jan. 25, 2008, at A1 (describing the 2008 tax rebate passed to boost the ailing economy).

[254] Again, this conclusion is not meant to imply that there is no scholarly consensus on the impact of taxes on labor supply. The notion that tax cuts increase revenue, for example, would be rejected by the vast majority of serious scholars. *See supra* notes 248, 249, and accompanying text (discussing different theories on the relationship between tax revenues and tax rates).

the behavior differences are much more likely to be caused by the differential tax rates rather than confounding factors. Take, for example, two individuals of similar educational backgrounds and work histories who are subject to different marginal tax rates. If the individual subject to a lower tax rate works many more hours than her counterpart subject to a higher tax rate, then this observation provides compelling evidence that high marginal tax rates significantly reduce labor supply. We therefore recommend a randomized experiment of marginal tax rates.

The unit of observation in this experiment should be the individual or household.[255] The critical outcome of interest in the tax debate is the impact of tax rates on labor supply and entrepreneurship. These decisions are made at the individual or household level, meaning that individuals or households are the appropriate units of observation.[256]

Imposing differential mandatory tax rates on similarly situated individuals might be controversial. One response to such potential debate would be to make it explicit that the government is sponsoring a lottery, the winners of which will receive a reduction in their tax rates. Only individuals who filed a tax return in the prior year (or perhaps only those who timely filed) might be deemed eligible for the lottery.[257] State-sponsored lotteries are common, and providing a prize for a fraction of those who meet a legal requirement might not seem objectionable. Even if only 0.01% of taxpayers were selected for the lottery, such a group would provide a relatively representative sample of over 10,000 taxpayers.

Alternatively, the government could randomly assign different mandatory marginal tax rates to individuals but then provide fixed lump sum transfers to those individuals who receive higher tax rates

---

[255] Note that by varying the unit of randomization between the individual and the household, policymakers can get a sense of the true effect of the "marriage penalty" and other important questions of tax policy. For more information on the marriage penalty, see generally James Alm et al., *Policy Watch: The Marriage Penalty*, 13 J. ECON. PERSP. 193 (1999).

[256] If policymakers want to study the spillover effects of taxes, such as whether benefits associated with lower taxes on the rich "trickle down" to the lower and middle classes, then policymakers can examine the behavior of each wealthy individual in greater detail. For example, if lower tax rates lead to greater entrepreneurship, then policymakers should examine the start-up businesses founded by those with lower tax rates and estimate the identities and salaries of employees of the start-up business. If this exercise proves impossible, then tax rates can be randomized at other units of observation, such as the state or county.

[257] We credit Terrence Chorvat for the idea of a state-sponsored lottery for individuals who meet tax law requirements.

so that average tax rates remain similar across individuals. There are several difficulties to this scheme, however. There will remain some differences in treatment, as the true average tax rate will depend on individual labor supply decisions, and these decisions will be differentially affected by different tax rates. In addition, an experiment that randomly assigns marginal tax rates *and* lump-sum transfers does not provide unambiguous estimates of the impact of different marginal tax rates. Instead, the experiment provides estimates of the effects of different marginal tax rates *and* offsetting transfers. If transfers also have an effect on labor supply—such as a wealth effect[258]—then the experiment fails in its aim to provide conclusive evidence about the impact of marginal tax rates on labor supply and entrepreneurship.

As with securities law experiments, a brief marginal tax rate experiment is unlikely to provide an unbiased estimate of the effect of different marginal tax rates.[259] If tax rates change for a brief time, individuals subject to low tax rates may shift work from future periods into the current period to take advantage of the lower tax rate. If people do act in this fashion, the experiment will generate an unrealistically high estimate of the impact of tax rates on labor supply. The experiment will reflect employees' abilities to shift work between time periods rather than to permanently adopt different labor arrangements in response to different incentives. A longer experiment period limits the ability of individuals to shift work between periods. One can easily move work from week to week, but it is much more difficult to move work from one year to another. As a result, the taxation experiment should take place over a relatively long period of time (such as two to three years), and outcome variables should be measured for at least a year after the conclusion of the experimental manipulation.

There are many outcome variables of interest for a randomized experimental study of different marginal tax rates. The most obvious outcome variables are labor supply and wages. The experiment we propose will directly address the degree to which lower taxes induce individuals to work more hours or seek more demanding, higher-wage jobs. Other outcome variables, such as entrepreneurship levels, child-

---

[258] *Cf.* Alan B. Krueger & Jörn-Steffen Pischke, *The Effect of Social Security on Labor Supply: A Cohort Analysis of the Notch Generation*, 10 J. LABOR ECON. 412, 434 (1992) (concluding that previous estimates of the wealth effect of Social Security—i.e., of the extent to which the program's benefits motivated people to exit the labor force—were exaggerated).

[259] *See supra* notes 240-41 and accompanying text (discussing why the experimental period needs to be relatively long).

care decisions, and unemployment rates, should also be examined. One of these outcome variables, or some weighted combination of them, might be selected as the target of a self-executing experiment, in which the result would be either slightly lower or slightly higher taxes for the population at large. This experiment may be particularly attractive if Democrats and Republicans on average have different empirical views about the effects of marginal tax rates. A self-executing experiment might leave each side optimistic that it will prevail, and it may be the only way to effect change if the losing side can be expected to conjure some rationalization for the outcome instead of changing its view on taxes.

3. Civil Rights

To this point, most of our examples of experiments have concerned corporate or public finance. But the idea of randomized testing can be applied to a larger set of laws that more directly regulate individual behavior. This subsection sketches how a randomized experiment could inform legislative choice concerning civil rights. At the moment, there is no federal law prohibiting employment discrimination on the basis of sexual orientation.[260] The Employment Non-Discrimination Act of 2007 (ENDA)—a minimalist prohibition of disparate treatment on the basis of sexual orientation—has been introduced in Congress several times,[261] and the House passed it in 2007,[262] but both chambers have yet to enact it. Even though polls suggest that an overwhelming majority of Americans oppose employment discrimination on the basis of sexual orientation,[263] opponents of ENDA

---

[260] Twenty-one states and the District of Columbia have passed state statutes that prohibit employers from discriminating on the basis of sexual orientation. *See Statewide Employment Laws and Policies*, HUMAN RIGHTS CAMPAIGN (July 26, 2010), http://www.hrc.org/documents/Employment_Laws_and_Policies.pdf (listing the states).

[261] *See, e.g.*, Employment Non-Discrimination Act of 2009, H.R. 3017, 111th Cong. (1st Sess. 2009).

[262] H.R. 3685, 110th Cong. (1st Sess. 2007).

[263] *See* THE GALLUP POLL: PUBLIC OPINION 2004 (Alec M. Gallup & Frank Newport eds., 2006) ("Americans overwhelmingly support 'equal rights in terms of job opportunities' for gay men and women."); Elizabeth Mehren, *Acceptance of Gays Rises Among New Generation*, L.A. TIMES, Apr. 11, 2004, at A1 ("72% favor laws to protect homosexuals from job discrimination . . . ."); John Newsome, *Employment Non-Discrimination Act (ENDA) Vote Tests Our Values: An Incrementalist Law Is a Blunder*, S.F. CHRON., Nov. 7, 2007, at B11 ("[A] 2006 Gallup Poll reveals that 89 percent of respondents favor equal opportunities for gay people."); GAY & LESBIAN ALLIANCE AGAINST DEFFAMATION, MEDIA REFERENCE GUIDE 29 (8th ed. 2010), *available at* http://www.glaad.org/document.doc?id=99 (noting that a 2005 Gallup poll showed 87% support for equal

argue that it would impose substantial litigation and other compliance costs on private employers.[264]

A 2000 General Accounting Office (GAO) study sheds some light on the question of litigation costs by analyzing the number of claims that had been made in the twelve states, including the District of Columbia, that had prohibited sexual orientation discrimination by private employers as a matter of state law.[265] One of us analyzed the claims data, together with more general employment data, and found that historically there has been only about one claim each year for every 60,000 workers.[266] If the employer's average cost per complaint were $100,000, the average annual cost of the statute per employee would be less than $2.[267]

While this analysis of historic data suggests that employer costs are low, these estimates might not fully represent the costs that a federal law would produce. For example, it is possible that employers in the first twelve states to pass the law are less likely to discriminate than those in the remaining thirty-eight. Or it might be possible that the specific language of ENDA would produce lower (or higher) costs of compliance than would similar state statutes. A randomized test of the impact of ENDA is a natural and powerful way to learn more about whether opponents' objections are well founded. A randomized control trial could produce valuable information on whether ENDA decreases the profitability or the stock price of firms. We would learn about the litigation and compliance costs for a representative subsample of firms. And we could even find out whether ENDA caused covered firms to lose market share to uncovered firms.

---

job opportunities for "homosexuals" but 90% support for equal job opportunities for—using different language—"gays and lesbians").

[264] *See, e.g.*, OFFICE OF MGMT. & BUDGET, EXEC. OFFICE OF THE PRESIDENT, STATEMENT OF ADMINISTRATION POLICY: H.R. 3685—THE EMPLOYMENT NON-DISCRIMINATION ACT (2007) (stating the policy position of former President George W. Bush, who opposed ENDA on "constitutional and policy grounds").

[265] *See* Letter from Barry R. Bedrick, Assoc. Gen. Counsel, U.S. Gen. Accounting Office, to the Honorable James M. Jeffords, Chairman, Comm. on Health, Educ., Labor and Pensions 2 (Apr. 28, 2000), *available at* http://archive.gao.gov/f0502/575711.pdf (finding "no indication that these laws . . . generated a significant amount of litigation"). The General Accounting Office became the Government Accountability Office in 2004. *Our Name*, U.S. GOV'T ACCOUNTABILITY OFFICE, http://www.gao.gov/about/namechange.html (last visited Jan. 15, 2011).

[266] Ian Ayres & Jennifer Gerarda Brown, *Mark(et)ing Nondiscrimination: Privatizing ENDA with a Certification Mark*, 104 MICH. L. REV. 1639, 1670 (2006).

[267] *Id.* at 1671.

In this subsection, we discuss how such a test might be structured. Although it would theoretically be possible to assign the application of ENDA randomly to individual workers, the administrative costs for an employer to comply with a discrimination prohibition on part of its workforce would not produce a very accurate view of firm-level costs of compliance. Thus, randomizing across firms would probably be the most effective approach. Given the negligible costs implicit in the GAO data, the compliance costs are unlikely to be so great as to create a substantial competitive disadvantage.[268] Firms assigned to the status quo control group (no prohibition of discrimination) might, however, be affected by the treatment group if employees are transferred to or from the treatment group because of the discrimination prohibition. If this type of overflow effect is large enough, it might militate for randomizing at the industry level—or conducting a mixed experiment that randomizes partially at the industry level and partially at the firm level.[269]

It is also necessary to determine what proportion of firms would be assigned to comply with ENDA. There are so many firms in the United States—more than six million businesses with employees[270]—that it would be possible to perform a powerful test that assigns perhaps one percent to the covered or uncovered arm of the experiment. The test might initially run for three to five years to give the firms and the employees time to learn about and adjust to the requirement.

A more libertarian version of the test would merely assign different ENDA defaults to different firms. Federal law currently allows employers to intentionally discriminate on the basis of employee sexual orientation. But this employer freedom to discriminate is nothing more than a default. There is nothing to stop employers from opting into ENDA by private contract and giving their employees and applicants virtually identical rights, including private rights of action, as they would have if ENDA had passed. Indeed, Ian Ayres and Jennifer Gerarda Brown have created a contractual mechanism where any employer can do just that

---

[268] By comparing relative market shares of the covered and uncovered firms, analysts can test for any impact on competition.

[269] Alternatively, the possible overflow effects of employees could be dampened by randomizing across cities or states. But the plausible size of this impact would not justify reducing the number of observations.

[270] U.S. CENSUS BUREAU, NUMBER OF FIRMS, NUMBER OF ESTABLISHMENTS, EMPLOYMENT, ANNUAL PAYROLL, AND ESTIMATED RECEIPTS BY ENTERPRISE EMPLOYMENT SIZE FOR THE UNITED STATES AND STATES, TOTALS: 2007, *available at* http://www2.census.gov/econ/susb/data/2007/us_state_totals_2007.xls (last visited Jan. 15, 2011) (reporting 6,049,655 U.S. firms in 2007).

with a few clicks at www.fairemploymentmark.org.[271] In this agreement, employers gain the right to use a certification mark if they promise not to discriminate on the basis of sexual orientation. The certification mark gives employers a private contract route to effectively opt into the statute's coverage. But Congress could take the fair employment idea further by giving firms an explicit right to affirmatively "opt into" ENDA coverage.[272]

The fight over civil rights legislation to date has exclusively sounded in terms of mandatory rules. But recent empirical research in behavioral economics suggests that defaults and menus matter, even at the firm-wide level.[273] Instead of running an experiment on the effects of mandatory ENDA, it is possible to test the impact of varying the default or menu dimensions of the law. Specifically, we could randomize firms into three groups: a control group with the status quo federal coverage, an "opt in" group of firms that can affirmatively opt for coverage by sending a notice to the Justice Department, and an "opt out" group of firms that can avoid liability under the statute by sending notice (in advance of any claimed discrimination) to the Justice Department that they do not wish to be covered.[274]

---

[271] Ayres and Brown advocated for the contractual mechanism of the Fair Employment Mark as a means to replicate ENDA in the private context. *See generally* Ian Ayres & Jennifer Gerarda Brown, *Privatizing Employment Protections*, 49 ARIZ. L. REV. 587 (2007). The fair employment license, however, falls short of ENDA protections on a few dimensions. *See* Ayres & Brown, *supra* note 266, at 1655 (noting that the license would not be enforced by governmental agencies, and private suits could not be brought in federal court).

[272] Ian Ayres, *Menus Matter*, 73 U. CHI. L. REV. 3, 8 (2006).

[273] *See* Yair Listokin, *What Do Corporate Default Rules and Menus Do? An Empirical Examination* 40 (Yale Law Sch. John M. Olin Ctr. for Studies in Law, Econ., & Pub. Policy, Research Paper No. 335, 2006), *available at* http://ssrn.com/abstract=924578 ("The presence or absence of corporate menus leads to large differences in outcomes, as do differences in default rules.").

[274] Randomized tests of default rules and menu options do pose particular problems for maintaining an uncontaminated control group similar to those described in Part III. It is possible that the treatment will impact the control group's behavior. For example, control-group firms may be confused about the legal regime under which they operate, or the existence of the treatment group might by itself increase the salience of the issue and put pressure on control-group firms to contract for substitutes for the treatment (such as the Fair Employment Mark). The availability of close substitutes for the treatment can bias (toward zero) the estimated impacts of the treatment. *See* James Heckman et al., *Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment*, 115 Q. J. ECON. 651, 655 (2000) ("When good substitutes for an evaluated program are available, the effect of the program will be small even if the effect of training is large.").

CONCLUSION

Randomized experimentation offers a powerful means to evaluate the effects of proposed policies. By applying laws and policies to different groups on a random basis, the causal impacts of the law can be isolated from other factors that would ordinarily be correlated with exposure to different policies. It is therefore not surprising that randomized controlled experiments have become increasingly prevalent in evaluating the impacts of different laws and policies. Legislators enact the vast majority of policy changes, however, without the benefit of randomized evaluations. This Article seeks to systematize and expand the use of randomized experiments in law and policy. In the short term, a number of individual experiments could advance the use of randomization and improve policy. In the long term, administrative agencies might be required to file randomization impact statements with all new regulations. Meanwhile, a norm in favor of experimental evidence can encourage legislators to back up their empirical claims with a willingness to initiate experiments through legislation, making policy outcomes dependent on experimental results.