

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. We explain why randomized testing is likely to produce better information than nonrandom evaluation of legal policies and 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 presumptively 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.

^{*} Professor of Law, George Washington University. The authors thank Anthony Vitarelli for excellent research assistance.

^{**} William K. Townsend Professor, Yale Law School.

^{***} Associate Professor, Yale Law School.

RANDOMIZING LAW

| | |
|---|----|
| I. INTRODUCTION | 3 |
| II. THE POWER OF RANDOMIZED CONTROLS | 6 |
| III. THE PROBLEMS OF NONRANDOM EVALUATION..... | 9 |
| A. Conventional Regression Analysis | 10 |
| 1. Omitted Variable Bias..... | 10 |
| 2. Publication Bias and Misspecification..... | 12 |
| B. The Laboratory of the States Reconsidered..... | 15 |
| IV. CAVEATS: LIMITS OF RANDOMIZATION STUDIES | 16 |
| A. Interpretive Problems..... | 16 |
| 1. Non-Double Blind Randomization | 17 |
| 2. Generalizability..... | 19 |
| a) Self-selection..... | 20 |
| b) Experimenter Selection..... | 21 |
| 3. Imperfect Randomization..... | 24 |
| a) Attrition..... | 24 |
| b) Crossover | 25 |
| c) Spillovers | 26 |
| B. Other Issues..... | 27 |
| 1. Costs..... | 27 |
| 2. Ethical Concerns | 29 |
| 3. Equality concerns..... | 32 |
| V. GUIDELINES AND APPLICATIONS..... | 38 |
| A. General Guidelines..... | 38 |
| B. Institution-Specific Guidelines | 42 |
| 1. Administrative Agencies: The Case for a Randomization Impact Statement.. | 43 |
| 2. Legislatures: The Case for Self Execution..... | 47 |
| C. Applications | 49 |
| 1. Securities Law..... | 49 |
| a) A Short Sale Experiment | 50 |
| b) Experimental Sarbanes-Oxley Repeal | 53 |
| 2. Tax Law | 57 |
| 3. Civil Rights | 60 |
| VI. CONCLUSION..... | 63 |

RANDOMIZING LAW

I. INTRODUCTION

Debates about the causal impacts of government policy are omnipresent. In 81 B.C., Chinese scholars debated 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 versus a competitive market. Over two thousand years later, theoretical debates over policies remain the norm. But many policy issues cannot be resolved by theory alone, 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 where the experimental population 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 1968,⁴ randomized “social experiments” on the impact of government policies. But the legal literature has virtually ignored them. Legal scholars have discussed the results of particular social experiments,⁵ and occasionally have commented that additional social experiments could

¹ *The Scholar as Government Consultant: The Great Salt and Iron Debate in Ancient China*, 7 AM. BEHAV. SCI. (1965), available at http://www.grazian-archive.com/archive/pdf/1965_06_00_ABS_salt_and_iron_debate_1_10.pdf.

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”); Alan Milner, *Restatement: The Failure of a Legal Experiment*, 20 U. PITT. L. REV. 795 (1959) (characterizing restatements of law as a failed experiment). The most prominent academic account of experimental approaches to government also defines experimentation broadly, mentioning randomization as a possible ingredient of experimentation only once. See Michael C. Dorf & Charles F. Sabel, *A Constitution of Democratic Experimentalism*, 98 COLUM. L. REV. 267, 348 (1998) (noting that systems for evaluating experiments “can themselves be benchmarked, and . . . can be combined with random-assignment experiments and other familiar methods of evaluation”).

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

⁴ A doctoral student, Heather Ross, developed the idea for the experiment, on the effect of a negative income tax, and then received governmental funding. The experimental results are reported in three volumes. See 1-3 THE NEW JERSEY INCOME MAINTENANCE EXPERIMENT (Harold W. Watts et al., 1976-1977). Useful summaries of the experiment are in SOCIAL EXPERIMENTATION 95-143 (Jerry A. Hausman & David A. Wise eds., 1985); and DAVID GREENBERG ET AL., SOCIAL EXPERIMENTATION AND PUBLIC POLICYMAKING 111-64 (2003).

⁵ See, e.g., Machaela M. Hoxtor, Comment, *Domestic Violence as a Crime Against the State: The Need for Mandatory Arrest in*

RANDOMIZING LAW

provide useful information in one field or another.⁶ But they have not addressed the normative question of whether the legal system should generally seek to incorporate experimental methods, and if so, what approaches the legal system should take to maximize the chance that experiments will improve policy.

Perhaps in part as a result of this scholarly neglect, past social experiments have clustered in specific policy areas. As the label “social experimentation” suggests, most of the experiments have been in the area of social services, testing whether expenditures on entitlements succeed in achieving social goals such as reducing poverty.⁷ For example, a recent experiment, executed under a Medicare statute requiring randomized testing of programs,⁸ assessed whether telephone contact by nurses to at-risk Medicare patients will reduce program costs.⁹ Another class of randomized studies test criminal justice policies.¹⁰ A rare exception outside these two areas was a set of experiments on electricity pricing.¹¹ Experiments almost never vary legal rights and obligations of ordinary citizens and entities in areas such as securities law or taxation.¹² Instead, they focus on possible provision of new services or on those who might be thought of as forfeiting rights by committing crimes.

This Article advances the case for randomizing law, including the legal rights and obligations expressed in statutes and regulation.¹³ Randomized experiments have the potential

California, 85 CAL. L. REV. 643, 655-57 (1997) (commenting on a Minneapolis experiment with randomized mandatory arrest of alleged domestic violence perpetrators).

⁶ See, e.g., Bernard E. Harcourt, *Post-Modern Meditations on Punishment: On the Limits of Reason and the Virtues of Randomization*, 74 SOC. RES. 307, 328-30 (2007) (proposing randomization in criminal justice, for example in setting the length of prison sentences); Laurens Walker, *Perfecting Federal Civil Rules: A Proposal for Restricted Field Experiments*, LAW & CONTEMP. PROBS., Summer 1988, at 67 (proposing randomized experiments on procedural rules).

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

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

⁹ See NANCY MCCALL ET AL., CENTERS FOR MEDICARE & MEDICAID SERVICES, EVALUATION OF PHASE 1 OF MEDICARE HEALTH SUPPORT (FORMERLY VOLUNTARY CHRONIC CARE IMPROVEMENT) PILOT PROGRAM UNDER TRADITIONAL FEE-FOR-SERVICE MEDICARE (June 2007); Reed Abelson, *Medicare Finds How Hard It Is To Save Money*, N.Y. TIMES, Apr. 7, 2008, at A1 (describing the program).

¹⁰ See generally David F. Farrington & Brandon C. Welsh, *A Half Century of Randomized Experiments on Crime and Justice*, 34 CRIMINAL JUSTICE 55 (2006) (providing an overview of randomized criminal justice experiments, the first of which was initiated in 1951 and reported on in 1978).

¹¹ See SOCIAL EXPERIMENTATION, *supra* note 4, at 11-53; see also RESEARCH TRIANGLE INST., ANALYTICAL MASTER PLAN FOR THE ANALYSIS OF THE DATA FROM THE ELECTRIC UTILITY RATE DEMONSTRATION PROJECTS (1978).

¹² For an exception that we propose to extend, see Part V.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

not merely to be governmentally funded academic exercises, but 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 what works. Just as random assignment of treatments is the most powerful method of testing for the causal impact of pharmaceuticals, randomly assigning individuals, firms, or jurisdictions to different legal rules can help resolve uncertainty about the consequences of laws and regulations.

Beyond endorsing randomized legal experimentation even 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, should accept decisions by agencies to randomize policies and perhaps even be more deferential to policy decisions arrived at after a process of experimentation. Ultimately, the executive branch could make formalized consideration of randomized control trials as central a part of the regulatory process as formalized consideration of the costs and benefits of regulations. If experimentation begins to occur sufficiently often in agencies, perhaps 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. In a self-executing experiment, an experiment can in effect serve to resolve a bet among competing legislative factions, with the experiment 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 II lays out the affirmative case for randomized control trials and describes our central proposal. Part III 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 IV 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. More

speculatively one might imagine courts themselves conducting prospective randomized control experiments to gather evidence on the most appropriate resolution in a case.

challenging problems from the perspective of policy implementation are that randomized legal policy may be costly or raise ethical concerns. Finally, Part V offers some guidelines for legal experimentation, including specific recommendations for legislatures and administrative agencies, and then describes some specific applications in which randomization seems especially likely to be fruitful.

II. 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 agricultural trials in research growing out of his work at the Rothamsted Experimental Station.¹⁴ In his 1935 book, *The Design of Experiments*,¹⁵ Fisher explained the power of the technique with the arresting example of a “lady [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.”¹⁶ 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 random order.”¹⁷

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:

In 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.¹⁸

In the term “randomized control trial,” it is the randomization itself that is producing the controlled environment of a similar comparison group. Of course randomization doesn’t mean that the control and treatment groups will be identical. If we looked at the heights of people in each group, we would see the standard bell curve. But the point is that we would see the same

¹⁴ RONALD A. FISHER, *STATISTICAL METHODS FOR RESEARCH WORKERS* (1925); see also IAN AYRES, *SUPER CRUNCHERS: WHY THINKING-BY-NUMBERS IS THE NEW WAY TO BE SMART* 46-80 (2007) (discussing power of randomization tool for business and NGOs).

¹⁵ RONALD A. FISHER, *THE DESIGN OF EXPERIMENTS* (1935).

¹⁶ RONALD A. FISHER, *THE DESIGN OF EXPERIMENTS* 11 (6th ed. 1951).

¹⁷ *Id.*

¹⁸ David P. Harrington, *The Randomized Clinical Trial*, 95 J. AM. STAT. ASSOC. 312 (2000).

bell curve of heights in both groups. The law of large numbers assures that in the limit 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 non-treatment dimension—random assignment creates groups that have statistically similar distributions on every non-treatment dimension. Since the distribution of height (or any other characteristic) is the same in both the control and the treatment groups, we can attribute any differences in the *average* group response to the difference in treatment.

Indeed Fisher’s breakthrough was in seeing that randomization could do a better job of producing a controlled experiment than would be possible with non-randomized controls. Fisher went so far as to argue that randomization produced better controls than could *ever* be achieved by physically matching the non-tested attributes. In discussing his “Lady and the Tea” problem, Fisher explained:

It is no sufficient remedy to insist that “all 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 by 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 on attributes that are unobservable to the researcher.

The first 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 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-

¹⁹ FISHER, *supra* note 16, at 18 (emphasis added).

²⁰ Medical Research Council, *Streptomycin in Tuberculosis Trials Committee: Streptomycin Treatment of Pulmonary Tuberculosis*, 2 BRIT. MED. J. 769 (1948).

RANDOMIZING LAW

controlled investigations, including clinical investigations, by experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved.”²¹ Since 1970, randomized clinical trials have been a critical part of FDA analysis of applications.²²

Given the considerable benefits of randomized policy experiments, we propose that 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 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, but a norm to randomize or explain why randomization could not be undertaken 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 Section. We discuss later the details of implementation (including when agency can proceed to regulate without an RIS).²⁴ The new norm, however, should be the presentation of data from a randomized policy experiment.

²¹ 21 U.S.C. § 355(d)(1) (2000); *see also* 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 (2007).

²² 21 C.F.R. § 314.50. *See* Abigail Alliance for Better Access to Developmental Drugs v. Von Eschenbach, 445F.3d 470 (D.C. Cir. 2006); Charles J. Walsh & Alissa Pyrich, *Rationalizing the Regulation of Prescription Drugs and Medical Devices: Perspectives on Private Certification and Tor Reform*, 48 RUTGERS L. REV. 883 (1996); *see also* 40 C.F.R. § 799.9420 (EPA regulation mandating randomized testing of toxic substances).

²³ The National Environmental Policy Act, 44 U.S.C. §§ 3501-21 (2000), requires an environmental impact statement (“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 impact.” *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, 58 Fed. Reg. 51,735 (Sept. 30, 1993). The objectives of this Executive Order are to enhance planning and coordination with respect to both new and existing regulations; to reaffirm the primacy of federal agencies in the regulatory decisionmaking process; to restore the integrity and legitimacy of regulatory review and oversight; and to make the process more accessible and open to the public.

²⁴ *See infra Part V.B.2.*

III. 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 the argument 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, 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 in 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 some changes may be attributed to them. But reasonable observers often disagree about causality. And even if reasonable sophisticated parties would agree, partisans may offer misleading interpretations of the data. The media may then summarize the debate by simply noting that experts disagree.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 associations between jurisdictions adopting policies and other variables need not imply causation. It will thus almost always be relatively easy for partisans to find some basis on which either to develop misleading results or to offer critiques of results that in fact are relatively robust. Part III explains why even with numerous jurisdictions, conventional multiple regression analysis in which the policy of interest forms an independent variable may produce inaccurate results. These sections, of course, are not intended to provide comprehensive overviews of the uses and limits of statistical analysis.²⁸ Part III.C comments on the difficulties of improving the law by using the states as policy laboratories without randomization.

²⁵ See *supra* note 2.

²⁶ See Bryan Keefer, *Tsunami*, COLUM. JOURNALISM REV., July 1, 2004, at 18 (discussing reporters’ reluctance to take sides on issues of public controversy).

²⁷ A similar problem exists when jurors try to assess evidence beyond their competence. See Scott Brewer, *Scientific Expert Testimony and Intellectual Due Process*, 107 YALE L.J. 1535 (1998).

²⁸ A useful overview of regression analysis is WILLIAM MENDENHALL & TERRY L. SINCICH, A SECOND COURSE IN STATISTICS:

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 could mean that sex education encourages students to have more sex, but it also could reflect that school districts with high rates of student sexual activity respond to these rates by offering sex education. A standard statistical approach to overcoming this problem is to add control variables for the characteristics of the students, such as family income, parents' education, and religion, as well as of the community, such as whether it is rural and in which region of the country it is located.³⁰ If those variables exhaust all nonrandom factors contributing to community and family decisions about sex education, then this technique will be successful, because the coefficient on the sex education variable then represents the effect of random variation in whether students are exposed to sex education. But if there is an omitted variable, correlated with both the community decision to offer sex education and the individual decision to have sex, the coefficient will be biased.

This problem cannot easily be avoided even by careful researchers (and can be exploited by researchers who hope to establish a particular result). There are at least two reasons for this. First, the available data may be incomplete. Even if there are strong theoretical reasons to believe that parental education is an important variable, it may be impossible to develop a measure that fully accounts for the parent's educational level.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 all high school graduates are alike, but within each group, there may be considerable educational heterogeneity. Even more precise data—including information like parental GPAs—will be at best only crude proxies. Second, the researchers' theoretical accounts of what variables may correlate with the dependent and independent variables are likely to be incomplete.

REGRESSION ANALYSIS (6th ed. 2003). For a critical analysis of the use of empirical evidence in legal scholarship, see Lee Epstein, *The Rules of Inference*, 69 U. CHI. L. REV. 1 (2002).

²⁹ See, e.g., Deborah Anne Dawson, *The Effects of Sex Education on Adolescent Behavior*, 18 FAMILY PLANNING PERSP. 162 (July/August 1986)

³⁰ See, e.g., *id.* at 170 tbl. 9 (listing control variables).

³¹ Dawson's study used a binary variable indicating whether the mother had at least twelve years of education. See *id.* at 166.

The omitted variable bias may be particularly problematic when regressions are used to analyze the behavior of individuals who have self-selected into particular governmental programs. For example, Julie Cullen et al. analyzed the effect of school choice lotteries, whose winners would be allowed to attend particular schools.³² Students who won the school choice lotteries tended to do better than students who did not enter the lotteries. Competing explanations include that lottery winners were allowed to attend better schools and that more motivated students are likely to self-select into the lottery. In the absence of variables fully capturing student motivation, a regression would tend to inflate the apparent effects of the schools on performance. Indeed, Cullen et al. showed that students who won the school choice lotteries performed no better than students who entered but lost the same lotteries. So it was student motivation, and not school quality, that caused the difference in performance between school lottery winners and non-lottery entrants. Though not created for the purpose of facilitating data analysis, the lottery produced random assignments that allowed the researchers to avoid omitted variable bias.

Even studies that attempt to control for all available information and seek to minimize the danger of omitted variable bias may nonetheless omit important variables. This can be shown by comparing the results of randomized experiments from the results of nonrandomized statistical analysis. Paul Glewwe et al. conducted separate prospective randomized and retrospective nonrandomized studies of whether making “flip charts” available to students in Kenya increased test scores. The retrospective studies showed that flip charts increased test scores, while the randomized studies revealed no effect.³³ Even a difference-in-difference analysis gave misleading results, showing that students in schools adopting flip charts performed especially well in flip chart subjects relative to other subjects. The forces that lead jurisdictions or 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. A full discussion of these

³² See, e.g., Julie Berry Cullen et al., *The Effect of School Choice on Student Outcomes: Evidence from Randomized Lotteries* (NBER Working Paper No. 10113, 2003).

³³ Paul Glewwe et al., *Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya* (NBER Working Paper No. 8018, 2000).

techniques is beyond our scope here, but these approaches will still often be 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,³⁴ 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. Casual empiricism, in any event, suggests that such studies require sufficient analytical judgment that their improved statistical power may not translate to a greater likelihood that the findings will be accepted in the public policy process. For example, a paper by Saurabh Bhargava and Vikram Pathania takes advantage of the discontinuity in cellular telephone rates around 9 p.m.³⁵ Call volumes today increase discontinuously around the 9 p.m. threshold, but there has been no increase in accidents immediately after 9 p.m. relative to the period before cell companies began to offer free calling after 9 p.m. Nonetheless, policymakers have continued to proclaim cell phone driving as dangerous as driving under the influence of alcohol.³⁶ Instrumental variables and regression discontinuity studies do not necessarily have greater impact on the policy process than other studies, even for 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. A finding of a statistically significant outcome, at the generally accepted 0.05 level, means that there is a five percent chance that an outcome at least as extreme would have occurred by pure chance if the null hypothesis were true.³⁷ If, for example, researchers test 100 propositions that in fact are all false and would be counterintuitive if true, about five of these tests may produce

³⁴ See Jerry A. Hausman, *Specification Tests in Econometrics*, 46 *ECONOMETRICA* 1251 (1978).

³⁵ Saurabh Bhargava & Vikram Pathania, *Driving Under the (Cellular) Influence: The Link Between Cell Phone Use and Vehicle Crashes* (AEI Working Paper No. 07-15, July 2007), at http://aei-brookings.org/admin/authorpdfs/redirect-safely.php?fname=../pdffiles/WP07-15_topost.pdf.

³⁶ See, e.g., Mike Stuckey, *Hands-Free Phones Are Lifesavers, Study Says*, MSNBC, May 13, 2008, at <http://www.msnbc.msn.com/id/24580099/> (quoting a California legislator who embraced a recent study on cell phones while apparently paying no heed to the Bhargava and Pathania study).

³⁷ An example of a “null hypothesis” would be that in the true relationship being estimated by a regression equation, the coefficient for one of the independent variables in fact is zero, indicating that, after controlling for other variables, there is no relationship between the dependent variable and that independent variable.

statistically significant results, and these mistaken results will be the most publishable.³⁸ Meanwhile, insignificant findings provide little support for the truth of the corresponding null hypotheses. Such findings also may be most publishable when they are counterintuitive, but a counterintuitive failure to reject a null hypothesis may also be the result of chance.³⁹

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,⁴⁰ which variables to include, what transformations to apply to the variables,⁴¹ and which observations to include.⁴² Especially within social science, researchers do not necessarily settle on regression specifications in advance, but instead “pretest” data to determine which results to report.⁴³ Considering a large number of regression specifications may help researchers develop nuanced accounts, but researchers will generally be more likely to report results producing statistical significance.⁴⁴ Laboratory experiments are also subject to publication bias, but other researchers can attempt replication. Social science researchers cannot rerun history.⁴⁵

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 deterrent effects of the death penalty.⁴⁶ For

³⁸ Some researchers have sought to counter this by encouraging the publication of statistically insignificant results. *See, e.g.,* Huai Yong Cheng, *The Potential Value of Negative Studies*, 6 J. AM. MED. DIRECTORS ASS’N 426 (2005).

³⁹ *See* J. Bradford De Long & Kevin Lang, *Are All Economic Hypotheses False?*, 100 J. POL. ECON. 1257 (1992) (conducting a statistical analysis of the distribution of statistical results to estimate the proportion of unrejected null hypotheses that are false). The De Long and Lang statistical analysis rejects “the null hypothesis that more than about one-third of *unrejected* null hypotheses ... are true.” *Id.* at 1258. That is, among published findings that do *not* show statistically significant outcomes

⁴⁰ *See, e.g.,* WILLIAM H. GREENE, *ECONOMETRIC ANALYSIS* 316-50 (3d ed. 1998) (providing an introduction to these issues).

⁴¹ *See id.* (considering the possibility of nonlinear specifications).

⁴² There may be flexibility both in determining the general coverage of the study (for example, what years or states to study), as well as in identifying outliers. Typically, when an observation is identified as an outlier, a researcher will run a regression both with and without the outliers to determine whether the results are robust. There are also econometric techniques designed to produce estimates not overly susceptible to outliers. *See, e.g.,* PETER J. ROUSSEEUW & ANNICK M. LEROY, *ROBUST REGRESSION AND OUTLIER DETECTION* (2003). Some researchers, however, may not use these techniques.

⁴³ 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.”).

⁴⁴ The traditional *t* statistic will be inaccurate when researchers test numerous regression specifications and then focus only on those whose *t* statistics appear to produce statistically significant results. *See, e.g.,* Dmitry Danilov & Jan R. Magnus, *Forecast Accuracy with Pretesting with an Application to the Stock Market*, 23 J. FORECASTING 251 (2004).

⁴⁵ *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”).

⁴⁶ *See* John J. Donohue & Justin Wolfers, *Uses and Abuses of Empirical Evidence in the Death Penalty Debate*, 58 STAN. L. REV. 791 (2005).

example, they criticized a study by Hashem Dezhbakhsh and Joanna Shepherd,⁴⁷ focusing first on a cross-sectional analysis of homicide trends in states that either abolished or adopted the death penalty. Dohonue and Wolfers argue that the same general trends existed in states that had not changed death penalty policy, and reanalyzed the data with a difference-in-differences approach. This produced statistically insignificant results.⁴⁸

This does not mean that every empirical question will yield an uncertain answer. But the death penalty is hardly the only debate about which scholarly experts have hotly contested empirical outcomes. Other recent examples in the criminological context include debates about whether abortion legalization is responsible for the decrease in the crime rate,⁴⁹ and whether statutes allowing citizens to carry concealed handguns lower violent crime rates.⁵⁰ Whatever the merits, academics and policymakers have not reached consensus on these questions. Even if the median voter or median decisionmaker would be swayed by empirical results, it will not be easy to determine what to believe.

Publication bias is a danger in randomized studies too.⁵¹ But there is less room for identifying alternative empirical specifications given the centrality of the random variable. As Esther Duflo has noted, in retrospective studies, “[e]x post the researchers or evaluators define their own comparison group, and thus may be able to pick a variety of plausible comparison groups,”⁵² but in a randomized study, the treatment and comparison groups will generally be clearly defined. There is still some danger that researchers will decide not to publish, but that danger is considerably reduced when governmental institutions have sponsored the research by supporting the randomization of policy and a particular set of researchers has promised to

⁴⁷ See Hashem Dezhbakhsh & Joanna M. Shepherd, *The Deterrent Effect of Capital Punishment: Evidence from a Judicial Experiment* (Emory Law & Econ. Research Paper No. 04-04, 2003), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=432621.

⁴⁸ Donohue & Wolfers, *supra* note 46, at 800-04.

⁴⁹ The paper that started the debate is John J. Donohue III & Steven D. Levitt, *The Impact of Legalized Abortion on Crime*, 116 Q.J. ECON. 379 (2001).

⁵⁰ At the center of this debate is the book JOHN R. LOTT, JR., *MORE GUNS, LESS CRIME* (1998).

⁵¹ Selective publication of results has been most clearly demonstrated in the medical arena, though the studies do not assess whether selective publication is a result of self-censorship by authors (perhaps because they do not want to suggest that a drug was ineffective) or by journals. See, e.g., Eric H. Turner, M.D. et al., *Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy*, 358 NEW ENG. J. MED. 252 (Jan. 17, 2008) (analyzing which reviews of antidepressant agents submitted to the FDA were subsequently published).

⁵² Esther Duflo, *Scaling Up and Evaluation*, 2004 ANN. WORLD BANK CONF. ON DEVEL. ECON. 341, 353.

analyze the effects of the experiment. Indeed, governments can virtually eliminate the risk by requiring publication of experimental results as a condition of funding.

B. The Laboratory of the States Reconsidered

For statistical research to influence policy, rather than merely serve as a sound bite, it must not only be executed well, but also be executed in a way that policymakers can understand and cannot ignore. These challenges pose hurdles for a frequent justification of federalism, that allowing states to make independent choices provides for a kind of laboratory testing of policy.⁵³ Susan Rose-Ackerman has shown that federalism may insufficiently promote experimentation for numerous reasons,⁵⁴ for example because one state may hope to free-ride on the activities of other governments.⁵⁵ Edward Rubin and Malcolm Feeley have similarly noted that experimentation sometimes may be expensive and likely not on balance beneficial for the experimenter,⁵⁶ and so centralized coordination may be needed to take full advantage of federalism.⁵⁷

Yet, at least sometimes, states do change their policies and take risks in doing so in the hope of achieving informational benefits.⁵⁸ As Barry Friedman notes, states may innovate for a variety of reasons, quite apart from any desire to engage in “experimentation.”⁵⁹ These state innovations serve at least a crude experimentation function.⁶⁰ Commentators may observe that

⁵³ The classic statement of this theory is Justice Brandeis’s. See *New State Ice Co. v. Liebmann*, 285 U.S. 262, 311 (1932) (Brandeis, J., dissenting) (“It is one of the happy incidents of the federal system that a single courageous state may, if its citizens choose, serve as a laboratory; and try novel social and economic experiments without risk to the rest of the country.”). For a discussion of this justification for federalism, Ann Althouse, *Vanguard States, Laggard States: Federalism and Constitutional Rights*, 152 U. PA. L. REV. 1745, 1750-76 (2004).

⁵⁴ See Susan Rose-Ackerman, *Risk Taking and Reelection: Does Federalism Promote Innovation?*, 9 J. LEGAL STUD. 593 (1980).

⁵⁵ *Id.* at 594. The possibility that there might be insufficient incentives to innovate is apparent even in areas in which state competition has generally been trumped, such as corporate governance law. See Michael Abramowicz, *Speeding up the Crawl to the Top*, 20 YALE J. ON REG. 139 (2003) (arguing that there are suboptimal incentives for states to innovate in corporate law). *But see* Roberta Romano, *The States as a Laboratory: Legal Innovation and State Competition for Corporate Charters*, 23 YALE J. ON REG. 209 (2006) (arguing that states are in fact effective laboratories in the corporate charter context).

⁵⁶ Edward L. Rubin & Malcolm Feeley, *Federalism: Some Notes on a National Neurosis*, 41 UCLA L. REV. 903, 923-26 (1994).

⁵⁷ *Id.* at 926 (noting that absent coordination by a central authority, “state-initiated experiments are unlikely to be truly useful to other states because of more specific, technical variations” among the states).

⁵⁸ See, e.g., *FERC v. Mississippi*, 456 U.S. 742, 787-88 (1982) (O’Connor, J., concurring in part and dissenting in part) (“[S]tate experimentation is no judicial myth.”).

⁵⁹ Barry Friedman, *Valuing Federalism*, 82 MINN. L. REV. 317, 399 (1997) (“‘Innovation’ might have been a better word choice for Justice Brandeis than ‘experimentation,’ saving us all a lot of bother.”).

⁶⁰ 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 how administrative agencies can serve as “the continuing organized link between the national and the local, helping to create through national action the local

one state's approach to a particular issue, such as health care reform,⁶¹ has gone particularly badly or well, and this may influence their decisionmaking. Federalism, however, does not easily facilitate a scientific approach to experimentation. The difficulty that social scientists and especially policymakers face in assessing the results of state innovations contributes to the inaptness of the states-as-laboratories metaphor.

Still, federalism may be more conducive to experimentation than alternatives. Previous commentators have noted that randomized experiments are much more common in North America than in the rest of the world,⁶² and speculated that federalism may help explain this.⁶³ In any event, the mere existence of different jurisdictions could be conducive to randomized experimentation in two ways. First, it may be possible to randomize policies across states, at least among states that consent. It would be more awkward to randomize policies in the absence of generally accepted jurisdictional boundaries. And second, states themselves can serve as loci for experimentation at smaller jurisdictional levels, such as cities and counties. Indeed, randomized experiments have increasingly been conducted within states.⁶⁴

IV. CAVEATS: LIMITS OF RANDOMIZATION STUDIES

A. *Interpretive Problems*

Advocates of randomized studies have emphasized that only this type of study can establish causality with high confidence. For example, Esther Duflo has argued that “while it is always possible to reject experimental results on the grounds that the experiment was poorly designed, or failed, when an experiment is correctly implemented (which is relatively easy to ascertain), there is no doubt that it gives us the effect of the manipulation that was implemented.”⁶⁵ But what “is relatively easy to ascertain” may still remain controversial in public debate. Moreover, even if the measured effects can be confidently traced to the “manipulation,” some extrapolation will generally be needed to assess the full consequences of a

conditions for experimentation, and changing national arrangements accordingly.”).

⁶¹ See, e.g., Sara Rosenbaum, *Mothers and Children Last: The Oregon Medicaid Experiment*, 18 AM. J.L. & MED. 97 (1992).

⁶² See GREENBERG ET AL., *supra* note 4, at 38 (noting as an exception that the Netherlands tested an unemployment-counseling program).

⁶³ *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.*

⁶⁴ See GREENBERG ET AL., *supra* note 4, at 37-38.

⁶⁵ Esther Duflo, *Field Experiments in Development Economics* (Jan. 2006), <http://econ-www.mit.edu/files/800>, at 23.

law enacting the legal experiment. This section explains that this is so because 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 of the possibility that randomization may be imperfect.

1. Non-Double Blind Randomization

The purest form of randomized experiments includes informational control on 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 researcher remains blinded about 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."⁶⁶ Keeping subjects in the dark as to whether they are in the treatment group or not 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 (from the Latin for "I will please") pill, often a sugar pill.⁶⁷

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 that they are subjects in an experiment. For example, suppose that randomly selected workplaces were to be 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. But the transparency of this randomization is not as large a concern here as in a medical context. Medical researchers are primarily interested in the impact of the drug independent of any psychological placebo effects. In the policy arena, researchers want to see how knowledge of the law impacts people's behavior. Information about whether a workplace is treated becomes part of the

⁶⁶ RON MCQUEEN & CHRISTINA KNUSSEN, INTRODUCTION TO RESEARCH METHODS AND STATISTICS IN PSYCHOLOGY (2006).

⁶⁷ Austin Flint in 1863 conducted the first placebo-controlled experiment when he treated a small number of hospital inmates for rheumatic fever. The control group received what Flint called a "placebo" or "placeboic remedy" of a "very largely diluted" tincture of quassia. See AUSTIN FLINT, A TREATISE ON THE PRINCIPALS AND PRACTICES OF MEDICINE (1866).

treatment, but this is not inherently bad, because the researcher wants to know whether a known legal change will have an impact.

There is, however, another problem. Even when subjects don't 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. In 1955 Henry Landsberger recognized, in a set of ergonomic experiments at the Hawthorne Works near Chicago, that subject knowledge of the experiment might affect subject behavior.⁶⁸ The researchers found a short-term improvement in worker performance after almost any change in lighting.⁶⁹ But productivity soon returned to normal levels. Although there remains some controversy over whether the experimental context did affect productivity in that experiment,⁷⁰ 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 in response to the substance of the experimental manipulation. Similarly, the phrase "John Henry effects"⁷¹ is used to describe changes in behavior in comparison groups that recognize that they are not being subject to experimental manipulations.⁷²

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 participate in the randomized trial.⁷³ In the legal context, sometimes knowledge of a change does not necessitate that subjects know that they are taking part in a randomized study. For example,

⁶⁸ HENRY A. LANDSBERGER, HAWTHORNE REVISITED (1958).

⁶⁹ RICHARD GILLESPIE, MANUFACTURING KNOWLEDGE: A HISTORY OF THE HAWTHORNE EXPERIMENTS (1985).

⁷⁰ Compare, e.g., WILLIAM H. WHYTE, JR., THE ORGANIZATION MAN 34 (1956) (claiming that increased productivity in an industrial experiment by Hawthorne was a result of experimenters' behavior toward those treated), with Stephen R.G. Jones, *Was There a Hawthorne Effect?*, 98 AM. J. SOCIOLOGY 451 (1992) (questioning the existence of the effect by scrutinizing Hawthorne's original study).

⁷¹ See, e.g., Allen C. Barrett & Doris A. White, *How John Henry Effects Confound the Measurement of Self-Esteem in Primary Prevention Programs for Drug Abuse in Middle Schools*, 36 J. ALCOHOL & DRUG EDUC. 87 (1991) (providing an alleged example of John Henry effects).

⁷² DUFLO ET AL., *supra* note 87, at 68 ("The comparison group may feel offended to be a comparison group and react by also altering their behavior (for example, teachers in the comparison group for an evaluation may 'compete' with the treatment teachers or, on the contrary, decide to slack off).").

⁷³ David A. Braunholtz, *Are Randomized Clinical Trials Good for Us (in the Short Term)? Evidence for a "Trial Effect,"* 54 J. CLINICAL EPIDEMIOLOGY 217 (2001).

one could imagine a test of speed limits where 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, at least the businesses subject to the new rules would be likely to find out the reason for the new rules, and it seems likely that others in the industry would also recognize the experimental context. This could, for example, 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 affect the ultimate policy result, though this is less likely to be a concern 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 section noted the difficulty of extrapolating from a sample with certain informational attributes—such as subjects knowing they were participating in an experiment—to a population with different informational attributes. There are analogous problems of extrapolation where 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, which “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.”⁷⁴ This may occur as a result of self-selection, because 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.

⁷⁴ James J. Heckman & Jeffrey A. Smith, *Assessing the Case for Social Experiments*, 9 J. ECON. PERSPECTIVES 85 (1995); see also James Heckman & Richard Robb, *Alternative Methods for Evaluating the Impact of Interventions*, in LONGITUDINAL ANALYSIS OF LABOR MARKET DATA 156-245 (James Heckman & Burton Singer eds., 1985); James Heckman & Jeffrey Smith, *Assessing the Case for Randomized Evaluation of Social Programs*, in MEASURING LABOUR MARKET MEASURES: EVALUATING THE EFFECTS OF ACTIVE LABOUR MARKET POLICIES 35-95 (Karsten Jensen & Per Kongshoj Madsen eds., 1993); James Heckman, *Randomization and Social Program Evaluation*, in EVALUATING WELFARE AND TRAINING PROGRAMS 201, 201-03 (Charles Manski & Irwin Garfinkel eds., 1992); James Heckman & V. J. Hotz, *Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training*, 84 J. AM. STAT. ASSOC. 862 (1989).

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 that is seeking exposure to an experimental policy. The causal impact of the experimental policy on this self-selecting group may be different than the causal impact of the policy on the average individual affected by the policy.⁷⁵ Chemotherapy drugs, for example, increase the life expectancy of some cancer patients, but decrease the life expectancy (because of their side effects) of those free of cancer. Volunteers for experiments on 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.

As with drugs, so with policies. The volunteers for a policy experiment only give an accurate estimate of the causal effects of the policy if the volunteers are representative of the group of individuals that will be affected by the fully enacted policy. Consider an experimental job skills program. People who volunteer for such a program may be particularly likely to be helped by this program. Experimenters can attempt to control for differential effects, but some of the variables that affect the response to the job skills program for volunteers 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 of the average low-skill person and experimenters cannot adjust their effect estimates to account for discipline. If policymakers are considering making the program mandatory for people of a certain skill level, then the experimental estimate of the program's effect using volunteers is biased. Volunteer experiments can, however, guide policymakers 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.

⁷⁵ 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 (2001).

Sometimes, treatment will 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 in general be more severe for legal experiments, because it often may be even more difficult in a legal context to control for other characteristics. We can make some headway in measuring the severity of cancer according to test results and symptoms. 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.

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 government can and has applied different rules and regulations to different individuals and businesses. Thus, for example, the Emergency Unemployment Compensation Act of 1991 authorized the U.S. Department of Labor to test the impact of a job search assistant program by randomly requiring certain recipients of unemployment insurance to participate in the program.⁷⁶

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 an experiment is tried in only one location or only with some nonrandom subset of the individuals and entities who would eventually be affected by a law. Cost considerations may justify such nonrepresentativeness, and indeed it is common for “demonstration projects” to be deployed in one or more particular regions rather than randomly.⁷⁷ At times, skepticism about inferences from an experiment on a nonrepresentative

⁷⁶ PAUL T. DECKER ET AL., ASSISTING UNEMPLOYMENT INSURANCE CLAIMANTS: THE LONG-TERM IMPACTS OF THE JOB SEARCH ASSISTANCE DEMONSTRATION (2000), available at <http://www.upjohninst.org/erdc/jsa/execsumm.pdf>; Marcus Stanley et al., *Developing Skills: What We Know About the Impacts of American Employment and Training Programs on Employment, Earnings, and Educational Outcomes* (Harvard Econ. Dep’t, Working Paper, 1998).

⁷⁷ For a discussion of the transition from local demonstration projects to projects on a national scale, see Duflo, *supra* note 52, at 342-45.

population may be justified.⁷⁸ For example, a randomized workplace safety experiment on a sample of small firms might not extrapolate easily to a sample of large firms. Sometimes, it may be feasible to conduct randomization at a national level, for example in choosing Medicare recipients who will receive extra follow-up phone calls. If policy is to be implemented at a national level, then this will provide a sound assessment of policy. Often, however, coordination and data-gathering needs may make this difficult.

Moreover, even if policy is to be implemented at a national level, that does not necessarily mean that a single uniform policy will be optimal. Randomized results give powerful and powerfully 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.⁷⁹ For example, imagine that a speed limit study randomizing across different cities shows that twenty mph limits produce *more* accidents than thirty mph limits. It might still be that small, rural cities fare better with the lower limit. It is possible to 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 sub-experiment. But because a population can be divided in any number of ways, and because 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.⁸⁰ Similar problems can affect randomized experiments. Suppose, for example, that the federal government tested the minimum wage by randomly selecting one percent of adults, allowing those selected the option of informing employers that they would not need to be paid minimum wage. In theory, eliminating the minimum wage might

⁷⁸ 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.”).

⁷⁹ James J. Heckman, *Detecting Discrimination*, 12 J. ECON. PERSPECTIVES 101 (1998).

⁸⁰ Duflo, *supra* note 65, at 21. This helps explain why researchers studying social norms through ultimatum games have experimented in developing countries, where it is feasible to make the stakes large enough to affect experimental subjects’ welfare. See, e.g., Robert Slonim & Alvin E. Roth, *Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic*, 66 ECONOMETRICA 569 (1998).

increase employment. But businesses may not think it worthwhile to change their hiring practices, or to risk dissension from inconsistent wages, to have the chance to hire a few workers at a lower wage. Data from such a study therefore might not reliably reflect the effect of eliminating the minimum wage.⁸¹

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.⁸² There is no guarantee that marginal effects will at least correctly identify the approximate impact of the policy. For example, in the hypothetical concealed carry experiment, a permanent law might encourage more people to possess concealed handguns than a temporary law, but it is not clear how the additional group of adopters differs from the group that responds even to the temporary law. Perhaps the initial responders will tend to include more criminals seeking to take advantage of the law and the subsequent group will be more law-abiding, but this is only speculation.

At other times, a temporary law may be a poor proxy for long term effects because the law will have dynamic as well as static effects. Studies, for example, that seek to assess private school choice plans may fail to capture what proponents of such plans claim is a principal benefit, that free enterprise will allow educational entrepreneurs to learn over time what works.⁸³ Other arguments, however, suggest that a static analysis might overestimate the benefits of free choice; for example, in the short term for-profit schools might be willing to lose money in the hope of increasing the chance of being permitted to continue to receive public funds in the future. As another example, critics of the time-of-use electricity experiments argued that with a longer term study, customers would buy appliances that would help them adjust their electricity use based on time of day.⁸⁴

⁸¹ This hypothetical study is not valueless, however. The fact that change is “not worth the trouble” suggests that the benefits of the experimental policy are limited to some degree.

⁸² See, e.g., Dean Karlan & Jonathan Zinman, *Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts* (June 25, 2007), available at http://ipa.phpwebhosting.com/images_ipa/ExpandingCreditAccess.v3.pdf (reporting an experiment in which lending criteria were loosened for randomly selected marginal borrowers).

⁸³ See, e.g., Terry M. Moe, *Beyond the Free Market: The Structure of School Choice*, 2008 BYU L. REV. 557, 571 (2008) (making this point).

⁸⁴ See, e.g., JOSKOW at 46 (noting that “these experiments only allow us to estimate short-run elasticities of demand, given

3. *Imperfect Randomization*

The above sections have addressed the danger that the tested population might differ systematically from the more general population to which a legal experiment would apply. There is also, however, another problem – ensuring that in the population experimented upon, 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 straightforward to ensure that the decisions made by the computer are followed and that the results are properly measured. Dangers include attrition (where some randomized individuals or entities drop out of a study), crossover (where some control group members receive the treatment, or vice versa), and spillovers (where the treatment has some effect on the control group as well).

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 improvements made by schools in a developing country. A school's randomization into a comparison group might increase drop-out rates.⁸⁵ If the drop-outs tend to be the weaker students, and if the measurement of school success depends on tests of current students, then the attrition produces an artificial hurdle for the treatment. Attrition can also bias results when randomization occurs at the level of the individual. In a medical study, for example, people who receive the treatment but then suffer severe side effects might refuse to participate further in the study, making those who continue with the treatment a nonrepresentative sample.

Given any degree of attrition, those reviewing a study may argue about the best interpretation of the results. Statisticians may attempt to impute measurements for those who drop out, by comparing their characteristics with those of other subjects.⁸⁶ But this solution is imperfect, because there might be some unmeasurable difference between those who continue in

existing appliance stocks”).

⁸⁵ See, e.g., Abhijit V. Banerjee et al., *Remedying Education: Evidence from Two Randomized Experiments in India*, 122 Q.J. ECON. 1235, 1245 (2007) (discussing this problem).

⁸⁶ See, e.g., Richard B. Miller & David W. Wright, *Detecting and Correcting Attrition Bias in Longitudinal Family Research*, 57 J. MARRIAGE & FAMILY 921, 923 (1995) (describing the standard method of incorporating a variable representing the probability of dropping out directly into the study).

an experiment and those who drop out. Ultimately, sound statistical judgment is needed to assess such models' reliability.

A more objective solution is to use matched samples.⁸⁷ If someone from the treatment group drops out, results of the corresponding match from the control group are not counted either. This approach can be used also when randomization is at the institutional or jurisdictional level, if individuals can be matched across institutions or jurisdictions. With matching, it is not necessary ex post to construct a model that seeks to correct for attrition bias, which would increase the danger of subjectivity or manipulation. Statisticians would be needed to assign the original matches based on observable characteristics, but the matching would be difficult to manipulate, before it is known who will drop out.

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, that is not easy to escape. But imperfections may occur nonetheless, especially if the government wishes to leave some room for later discretion. For example, crossover can also occur if well-connected people can thwart random assignment. Alan Krueger, studying the effect of student to teacher ratios, found that some parents had managed to convince schools to reallocate their children from large to small classes.⁸⁸ This dilutes the treatment, as the small classes become larger, and means that the treated students on average will come from relatively highly motivated families.

Once again, statistical correctives exist. Under an “intention-to-treat” methodology,⁸⁹ an individual or entity who crosses over is counted with the group to which that person or entity was originally assigned. This reduces the measured effect of the treatment. Statisticians can compensate for the bias introduced by the intention-to-treat approach with a simple mathematical

⁸⁷ See ESTHER DUFLO ET AL., USING RANDOMIZATION IN DEVELOPMENT ECONOMICS RESEARCH: A TOOLKIT 35-36 (2006), <http://www.povertyactionlab.com/papers/Using%20Randomization%20in%20Development%20Economics.pdf> (“An extreme version of blocked design is the pairwise matched design where pairs of units are constituted, and in each pair, one unit is randomly assigned to the treatment and one unit is randomly assigned to the control.”).

⁸⁸ Alan B. Krueger, *Experimental Estimates of Education Production Functions*, 114 Q.J. ECON. 497, 505 (1999) (reporting higher attrition rates of students in smaller classes).

⁸⁹ See, e.g., Guido Imbens & Joshua Angrist, *Identification and Estimation of Local Average Treatment Effects*, 62 ECONOMETRICA 467 (1994) (discussing this approach)..

formula.⁹⁰ Assuming that it is possible to measure who ended up receiving the treatment and who ended up receiving the control, the formula can be applied mechanically, without allowing any discretion to the experimenters, and will generally improve the estimate of the treatment effect. This is an imperfect adjustment, because those who cross over may differ systematically from those who do not. Once again, this illustrates that even with randomized statistical methodologies, such statistical judgment may be needed to interpret the study results.

c) *Spillovers*

The final danger is that the treatment will spill over on the control group. Suppose, for example, that a shame sanction reduces recidivism not only in those who are shamed, but also in those who are randomized to the control group but hear about the shaming. Or, suppose that firms randomized to a relaxed securities disclosure regime decide that they want to disclose as much as their competitors. The comparison of treatment and control groups will underestimate the effects of the intervention. On the other hand, suppose that a random experiment eliminates speed limits on a random set of roads. Some drivers on the control roads may conclude that police, needing to fill their time somehow, will devote extra attention to the control roads. If these drivers slow down, measurements of the speed differential will be exaggerated.

A sometimes feasible solution is to randomize across geographical areas rather than individuals. Edward Miguel and Michael Kremer showed that randomized studies at an individual level had underestimated the benefits of deworming drugs, which benefited those in the immediate area who had not received the drugs.⁹¹ Randomizing across geographical areas largely solved the problem. This solution is not without drawbacks, however. Especially if the number of jurisdictions is small, a comparison of changes in treatment and control jurisdictions may not have much statistical power. In addition, some people may move to take advantage of the law elsewhere.⁹²

⁹⁰ 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 52, at 354.

⁹¹ Edward Miguel & Michael Kremer, *Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities*, 72 *ECONOMETRICA* 159 (2004).

⁹² Randomization across geographic areas can produce Tiebout sorting in much the same way as endogenous policy selection. See generally Charles Tiebout, *A Pure Theory of Local Expenditures*, 64 *J. POL. ECON.* 416 (1956) (providing the seminal account of the effects of citizen mobility).

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 an expensive task. 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 creates several complications. First, policymakers must determine means to measure the outcomes of interest. At times, the outcomes of interest may be reflected in preexisting data collection efforts, 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 non-compliance problem. Individuals are not mice and may find ways to avoid "complying" with the experimental treatment. Policymakers must find legitimate means of limiting such non-compliance, but such means will generally be costly. Even so, there will always be some number of non-compliers, and policymakers must ascertain means of preventing attrition and non-compliance from biasing the results of the experiment.

It is possible to use randomized testing to test many different variations on policies. For example, a test of speed limits could allow for a wide range of speed limits, or could test whether a tailored policy (of thirty mph in small cities and twenty mph in large cities) is superior to an untailored policy (of thirty mph in all cities).⁹³ But the possibilities for tailoring in any particular

⁹³ Randomized testing of this kind on the Internet has shown, for example, that tailoring a retail website's landing pages to be contingent on specific search queries produces more sales than a one-size-fits-all homepage. Thus, clicking on a Google ad for www.musiciansfriend.com after searching for "electric guitar" will take you to a different page than clicking on the same ad does after searching for "electric bass" because randomized testing of contingent strategy by Omniture showed higher revenue per customer of 15% when the landing pages were tailored to the specific search queries. Conversation with Matt Roche, President, Omniture, (June 14, 2007).

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 the choice of scarce options to test with full awareness that untested policies may still dominate. Sometimes, it may be worthwhile to focus on what seems to be the most attractive possibility, even if there is some chance that later a more attractive option will emerge.

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 reduce the costs of explaining the experimental policy to the subjects of the policy. Thus, the marginal costs of experimental policies should be declining 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 reduce the marginal costs of experimentation, but cannot eliminate them. As a result, policymakers should first pursue experiments of policies that have low experimentation costs, all else equal. While it is impossible to provide a complete description of 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.⁹⁴ In other words, policymakers should experiment with the best candidates first. This will reduce the direct costs of experimentation on the subjects of the experimental policy. Meanwhile, experiments should generally be as modest as possible but 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 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 of the

⁹⁴ Yair Listokin, *Learning Through Policy Variation*, 118 YALE L.J. 480 (2008) argues that, in many cases, the expected effect of policy is less important than the variance of the expected effects. Other things equal, however, higher expected value policies are superior to lower expected value policies.

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 a resolution of whether the government must obtain informed consent. Even with an informed consent requirement, randomized experimentation could still occur for many policies. For example, there will generally be no ethical objections to an experiment like the Medicare experiment,⁹⁵ where any participant may choose not to receive the services offered by the government. (There may be objections based on inequality among those who volunteer for the experiment, an issue to which the Article will return below.⁹⁶) Second, an existing collection of essays already explores this issue in considerable detail.⁹⁷

This section will briefly summarize and develop the argument that legal experimentation imposes no ethical hurdles beyond those inherent in general legal policymaking, while also sketching the opposing position. The argument against an informed consent requirement distinguishes legal from medical experimentation,⁹⁸ where informed consent is generally required.⁹⁹ An unconsented-to medical treatment violates a patient's bodily integrity rights.¹⁰⁰ The problem is not randomization. A state could not insist that all of its citizens take a new drug.

⁹⁵ See *supra* note 9 and accompanying text.

⁹⁶ See *infra* Part V.B.1.b.

⁹⁷ See ETHICAL AND LEGAL ISSUES OF SOCIAL EXPERIMENTATION (Alice M. Rivlin & P. Michael Timpane eds., 1975).

⁹⁸ 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.

Alice M. Rivlin & P. Michael Timpane, *Introduction and Summary*, in ETHICAL AND LEGAL ISSUES OF SOCIAL EXPERIMENTATION, *supra* note 97, at 1, 5.

⁹⁹ See Kathryn A. Tuthill, *Human Experimentation*, 18 J. LEGAL MED. 221 (1997).

¹⁰⁰ An early legal case insisting on informed consent frames the problem in these terms: "Every human being of adult years and sound mind has a right to determine what shall be done with his own body; and a surgeon who performs an operation without his patient's consent commits an assault, for which he is liable." *Schloendorff v. New York Hosp.*, 105 N.E. 92, 93 (N.Y. 1914).

Any rights that the individual has against the state constrain the state in legal experimentation. So, if a person has a right not to have property taken by the state,¹⁰¹ then the state cannot take that property in an experiment. But to the extent that a government can enact a policy generally, on the Lockean theory of implicit consent,¹⁰² there should be no ethical bar to the state enacting the policy only against a random set of individuals.

The opposing position flows from the Kantian principle that each person should be treated as an end rather than only as a means.¹⁰³ This principle also does not uniquely condemn randomization. Suppose a jurisdiction decides to enact a new universally applicable policy, even though policymakers suspect that it will not be effective but has enough of a chance of success to make it worth trying. If this counts as treating people as means only, then the ethical permissibility of a new policy must be judged excluding from consideration any benefit from the fact that implementation of the policy will produce information about the policy. But many legal regimes, such as patent law and securities law, are justified in part on the basis that they improve information. Information produced by a policy about the policy itself should not be uniquely condemned to irrelevance.

But assuming then that experimentation with universally applicable policies is ethical, is *random* policy experimentation ethical as well? An affirmative case focuses on the benefit of randomization, that it will produce better information than nonrandomized experiments.¹⁰⁴ Although this may at first appear to be a purely consequentialist justification, Robert Veatch argues that subjects of research have a right not to be put “at risk in an unnecessary experiment or one inefficiently designed.”¹⁰⁵ The Nuremberg principles on medical experimentation emphasized the importance of experimental design.¹⁰⁶ If universal experimentation is permissible, there is an *a fortiori* argument that random experimentation must be permissible as

¹⁰¹ See, e.g., U.S. CONST. amend. V (“[N]or shall private property be taken for public use, without just compensation.”).

¹⁰² See Peter G. Brown, *Informed Consent in Social Experimentation: Some Cautionary Notes*, in ETHICAL AND LEGAL ISSUES OF SOCIAL EXPERIMENTATION, *supra* note 97, at 95-96.

¹⁰³ See IMMANUEL KANT, GROUNDWORK OF THE METAPHYSIC OF MORALS 429 (H.J. Paton trans., Harper & Row 1964) (1785) (“Act in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means but always at the same time as an end.”).

¹⁰⁴ See *supra* Part IV.

¹⁰⁵ Robert M. Veatch, *Ethical Principles in Medical Experimentation*, in ETHICAL AND LEGAL ISSUES OF SOCIAL EXPERIMENTATION, *supra* note 97, at 21, 37.

¹⁰⁶ *Id.* at 37-38.

well. The difference between the universal and the random experiment is that some people do *not* receive the treatment. Unless there is an equality right to receive the treatment,¹⁰⁷ this difference should not make the experiment more problematic.

Medical experimentation itself further supports the argument that if universal policy experiments are permissible, policy experiments must be permissible as well, because medical experiments can generally be viewed equivalently as legal experiments. Subjects in medical experiments who give informed consent presumably would prefer a guarantee of receiving the treatment rather than a chance of a placebo. The status quo is a legal regime that constrains liberty by forbidding distribution of the treatment. Let us assume that the legal prohibition on what Eugene Volokh has called “medical self-defense” is permissible.¹⁰⁸ When the government authorizes a medical experiment,¹⁰⁹ it is effectively authorizing a new legal regime in which patients are permitted to have access to a treatment. The government does not authorize this new legal regime in a universally applicable way, but instead insists on randomization. Only some patients will legally have access to the treatment. It is thus sometimes permissible for new legal policies, including potentially pernicious ones, to be introduced randomly.

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

Those who defend the legitimacy of medical experimentation must either develop an account of which baselines are permissible or allow legal policymakers to play the same game

¹⁰⁷ See *infra* Part V.B.1.b.

¹⁰⁸ See generally *Abigail Alliance for Better Access to Devel. Drugs v. Eschenbach*, 495 F.3d 695 (D.C. Cir. 2007) (en banc) (finding no constitutional right to investigational drugs); Eugene Volokh, *Medical Self-Defense, Prohibited Experimental Therapies, and Payment for Organs*, 120 HARV. L. REV. 1813 (2007) (arguing for a right to medical self-defense).

¹⁰⁹ Government experimentation is necessary for at least some medical experiments. In the United States, the FDA reviews small-scale Phase II trials to determine whether to permit large-scale Phase III trials. See, e.g., Michael D. Greenberg, *AIDS, Experimental Drug Approval, and the FDA New Drug Screening Process*, 3 NYU J. LEGIS. & PUB. POL’Y 295, 305 (1999-2000).

outside the medical context. Existing randomized legal experiments generally allow opt-in to an apparently more favorable treatment. For example, drug offenders may receive the option of participating in an experiment in which they might be randomized to a drug court.¹¹⁰ This makes experimentation a one-way ratchet, allowing testing within an existing draconian regime of a more lenient alternative, but not allowing testing of more draconian legal approaches. The only way to test the more draconian approaches would be to change the baseline to those approaches, and then to allow individuals to opt into an experiment in which they might receive more lenient treatment. Similarly, one could test raising the speed limit (by allowing drivers to opt into a program in which they are permitted to drive 10 mph over the limit), but to test lowering the speed limit, policymakers must change the baseline. An insistence on informed consent privileges the status quo legal regime over alternatives, even if neither the status quo nor the alternative applied universally violates any rights.

3. *Equality concerns*

Concerns about informed consent focus on the rights of those subject to the experiment. Concerns about equality focus on the rights of those who either are randomly excluded from an experiment or who are assigned to the less desirable of the treatment and control groups. The equality concern is not limited to random experimentation, but extends also to cases in which a government with limited resources distributes those resources at random.¹¹¹ For example, governments have used lotteries to distribute scarce low-income housing,¹¹² rights to immigrate,¹¹³ and positions in magnet and charter schools.¹¹⁴ Maurice Rosenberg points out that random experimentation may be inevitably in tension with the “equal protection principle . . . that persons subjected to significantly different treatments must be shown to be different in ways

¹¹⁰ See, e.g., Denise C. Gottfredson & M. Lyn Exum, *The Baltimore City Drug Treatment Court: One-Year Results From a Randomized Study*, 39 J. RES. CRIME & DELINQ. 337 (2002)

¹¹¹ Such distribution has generally raised fewer objections than randomization for experimental purposes alone, and as a result experimentation has been particularly feasible in cases in which arbitrary decisions needed to be made in any event. See GREENBERG ET AL., *supra* note 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”).

¹¹² See, e.g., Denise Irene Arnold, *Lottery Prize Is Affordable Homes*, N.Y. TIMES, Feb. 7, 1988, at 12 (discussing a local housing lottery).

¹¹³ See, e.g., 8 U.S.C. § 1153(e)(2) (providing for distribution of some visas “strictly in a random order”).

¹¹⁴ See, e.g., Cullen et al., *supra* note 32 (analyzing such a lottery).

that supply a reasonable basis for the differences in treatment.”¹¹⁵ If equal protection were interpreted to prohibit all arbitrary legal differences among similarly situated individuals, then both random experimentation and other programs using random selection to award scarce resources must be eliminated.

There are, however, advantages to using randomization in both these contexts. In the experimental context, randomization has benefits already discussed,¹¹⁶ 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.¹¹⁷ 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. So, in any event, concludes Judge Friendly’s opinion in the leading case on this issue.¹¹⁸ 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.”¹¹⁹ A court someday might fail to follow or even overturn this precedent, but it reinforces the plausibility of the legal argument that randomization does not violate the Equal Protection Clause.¹²⁰

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.”¹²¹ “Why should Parliament,” he asks, “not make abortion criminal for pregnant women who were born in even years but not for those born in odd

¹¹⁵ Maurice Rosenberg, *The Impact of Procedure-Impact Studies in the Administration of Justice*, LAW & CONTEMP. PROBS., Summer 1988, at 13, 17.

¹¹⁶ See *supra* Part IV.

¹¹⁷ Rent-seeking can still occur if large numbers of individuals may 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 (1993) (analyzing a government lottery that produced 320,000 applications).

¹¹⁸ *Aguayo v. Richardson*, 473 F.2d 1090, 1108-10 (2d Cir. 1973).

¹¹⁹ *Id.* at 1109-10. One commentator has criticized the court for not indicating that its decision would be valid only for as long as the experimental program’s value were uncertain. Note, *Reforming the One Step at a Time Justification in Equal Protection Cases*, 90 YALE L.J. 1777, 1783 (1981).

¹²⁰ Randomization schemes may sometimes violate other constitutional provisions, however. See, e.g., *Delaware v. Prouse*, 440 U.S. 648 (1979) (finding random stops of vehicles to check driver’s license and registration inconsistent with the Fourth Amendment).

¹²¹ RONALD DWORKIN, LAW’S EMPIRE 178-84 (1986).

ones?”¹²² 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.¹²³ Dworkin claims that checkerboard statutes offend a principle that he calls “integrity.”¹²⁴ 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.”¹²⁵ That does not occur with random experimentation, where a single principle, the need to obtain more information, justifies both the treatment and control conditions.¹²⁶

Dworkin’s concern is that randomness seems arbitrary, but arbitrariness is often more troubling when it is non-random. Consider, for example, the different approaches of Justice Stewart and Justice Marshall in *Furman v. Georgia*¹²⁷ 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.”¹²⁸ 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.”¹²⁹ If Marshall was correct (and there is abundant evidence that he was)¹³⁰ that the death penalty is disproportionately visited upon the poor, the ignorant, and the underprivileged, then Justice Stewart cannot be right that the death sentence is randomly assigned. Marshall’s concern resonates with ex-ante equal protection concerns,¹³¹

¹²² *Id.* at 178.

¹²³ *See id.* at 185 (relating the checkerboard statute issue to conceptions of equality).

¹²⁴ *Id.* at 183.

¹²⁵ *Id.* at 184.

¹²⁶ Another example of Dworkin’s reaffirms that arbitrary distinctions are acceptable where 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.

¹²⁷ 408 U.S. 238 (1972).

¹²⁸ *Id.* at 313; *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.”).

¹²⁹ *Id.* at 365-66 (Marshall, J., concurring).

¹³⁰ A. DAVID C. BALDUS, ET AL., *EQUAL JUSTICE AND THE DEATH PENALTY: A LEGAL AND EMPIRICAL ANALYSIS* (1990).

¹³¹ *Stevens v. Marks*, 383 U.S. 234 (1966).

because citizens are treated differently from the get-go because of arbitrary characteristics. 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*,¹³² expressed a concern with a hypothetical clemency procedure: “[I]t is not too difficult to imagine extreme situations in which federal due process would be offended. For example, a procedure in which a governor or parole board merely pulled names out of a lottery bin or flipped coins to make clemency decisions would undoubtedly constitute a ‘meaningless ritual.’”¹³³ Language of this kind suggests that courts might be hostile to truly random application of law. The New York State Commission on Judicial Conduct in 1982 removed Jeffrey Jones, a Manhattan Criminal Court judge, from office for deciding in open court between a twenty- and thirty-day criminal sentence on the basis of a coin flip.¹³⁴ More recently, the Virginia Supreme Court removed trial judge James Michael Hull from office for determining parental custody rights for a Christmas holiday by flipping a coin.¹³⁵ The Supreme Court rejected Judge Hull's rationale that the probabilistic decision was an attempt to encourage the parents to resolve the dispute for themselves.¹³⁶ Federal Judge Gregory Presnell similarly used randomization as “a new form of alternative dispute resolution” when he ordered two attorneys to resolve a deposition dispute by playing a game of rock, paper, scissors.¹³⁷

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

¹³² 523 U.S. 272, 288 (1998).

¹³³ *Id.* at 288; *see also id.* at 288 (“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 . . .”).

¹³⁴ *People v. Jones* (N.Y. Crim. Ct. 1982).

¹³⁵ Gary Slapper, *Weird Cases: Justice by Coin-Toss*, TIMES ONLINE, Nov. 16, 2007, <http://business.timesonline.co.uk/tol/business/law/article2882090.ece>.

¹³⁶ *Id.*

¹³⁷ Adam Liptak, *Lawyers Won't End Squabble, so Judge Turns to Child's Play*, N.Y. TIMES, June 9, 2006, *available at* http://www.nytimes.com/2006/06/09/us/09judge.html?_r=1&oref=slogin; Jeralyn Merritt, *The “Rock, Paper Scissors” Judge*, TALKLEFT, June 9, 2006, *available at* <http://www.talkleft.com/story/2006/06/09/305/45461>.

are a variety of public policy rationales for randomized decisions. It's not clear whether Judge Hull was sincere in claiming that his coin flipping over Christmas 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.¹³⁸ Disputants bargaining in the shadow of probabilistically divided, Solomonic rights have powerful incentives to speak more honestly with each other—and therefore may be more likely to settle a dispute before the actual coin flip,¹³⁹ just as the lawyers in the deposition dispute resolved their dispute before having to play rock, paper, scissors on the courthouse steps. Moreover, in the context of child custody, Jon Elster has proffered an independent rationale for resolving custody disputes by coin flipping.¹⁴⁰ Elster argues that probabilistically assigning custody in close cases is valuable because the state does not tell the child that one parent is marginally better than the other. For Elster, publicly stating that mom or dad is the marginally better Christmas custodian may not be in the best interest of the children.

Judicial antipathy to randomized decisions is at its highest with regard to decision-making 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 certain sentences.¹⁴¹ New York State might get a bigger bang for its incarceration buck if it followed Judge Jones and flipped coins for twenty- and thirty-day sentences instead of sentencing everyone to twenty-five days.¹⁴² (This deterrence result is, however, reversed for risk-preferring criminals, and it is

¹³⁸ Ian Ayres & Eric Talley, *Solomonic Bargaining: Dividing a Legal Entitlement to Facilitate Coasean Trade*, 104 YALE L.J. 1027 (1995).

¹³⁹ Solomonic entitlements have an “information forcing” effect on ex-ante bargaining because disputants no longer are 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 a plaintiff to enter into two different kinds of settlement—one where he or she buys the defendant’s probabilistic entitlement and one where he or she sells his or her own probabilistic entitlement. The offsetting incentives to overstate value as a seller and understate value as a buyer lead to more forthright and efficient negotiations. *Id.*; see also PETER CRAMTON, ROBERT GIBBONS & PAUL KLEMPERER, *DISSOLVING A PARTNERSHIP EFFICIENTLY* (1987).

¹⁴⁰ JON ELSTER, *SOLOMONIC JUDGMENTS: STUDIES IN THE LIMITATION OF RATIONALITY* (1989).

¹⁴¹ Steven Shavell, *Economic Analysis of Public Law Enforcement and Criminal Law* (Nat’l Bureau of Econ. Research, Working Paper No. 9698, 2003).

¹⁴² Cf. David Lewis, *The Punishment That Leaves Something to Chance*, 18 PHIL. & PUB. AFF. 53, 58-62 (1999) (defending punishments where the severity is randomized, en route to justifying harsher penalties for those who by coincidence cause more harm).

troubling that Judge Jones before flipping did not inquire of the defendant if he was a betting man.)¹⁴³ But to our minds an even stronger rationale for randomization—even with regard to criminal sentencing—is to learn. After centuries of experience, we still do not have definitive evidence on whether longer sentences rehabilitate or harden criminals.¹⁴⁴ Justice O’Connor is appalled by the idea of clemency by chance, and randomness applied in a single case seems unlikely to produce useful information. There might, however, be value in randomly granting clemency and parole to inmates selected at random, to see if in fact they had a higher recidivism rate than those who were not selected.

The informational rationale for randomization also acts as a principle for deciding when not to test and when to stop testing. We shouldn’t allow randomized tests of parachutes,¹⁴⁵ 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.¹⁴⁶ 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 convenient analytically 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.¹⁴⁷ 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.¹⁴⁸ Notwithstanding the supposed requirements of informed consent, medical trials routinely fail to give participants the best current information about the likely result of the trial.¹⁴⁹ The reason for the failure is keep patients participating. Patient surveys indicate, unsurprisingly, that “[w]illingness to

¹⁴³ Editorial, *For Whom the Coin is Tossed*, N.Y. TIMES, Feb. 13, 1982, at 24.

¹⁴⁴ Jeffrey R. Kling et al., *Experimental Analysis of Neighborhood Effects*, 73 ECONOMETRICA 83 (2007).

¹⁴⁵ See *supra* note 7.

¹⁴⁶ Sarah J.L. Edwards, R.J. Lilford & J. Hewison, *The Ethics of Randomised Controlled Trials from the Perspectives of Patients, the Public, and Health Care Professionals*, 317 BMJ 1209-12 (1998).

¹⁴⁷ Moreover, from an efficiency perspective, it is sometimes cost effective to test and eliminate low probability therapies that might teach us a lot. *Supra* note 43; Martin L. Weitzman, *Optimal Search for the Best Alternative*, 47 ECONOMETRICA 641 (1979).

¹⁴⁸ R.J. Lilford & J. Jackson, *Equipoise and the Ethics of Randomisation*, 88 J. ROYAL SOC’Y MED. 552 (1995).

¹⁴⁹ Sarah J.L. Edwards, R.J. Lilford & J. Hewison, *The Ethics of Randomised Controlled Trials from the Perspectives of Patients, the Public, and Healthcare Professionals*, 317 BMJ 1209 (1998) (“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.”).

RANDOMIZING LAW

undergo randomisation drops as prospective participants are given more preliminary data and as they are made aware of any accumulating evidence of effectiveness.”¹⁵⁰

V. GUIDELINES AND APPLICATIONS

We saw in Part III that even with the best statistical tools, it is often difficult to make inferences about causality from nonrandomized policy changes. Randomization makes interpretation much easier, though as we saw in Part IV, even randomized experiments can be difficult to interpret. Given these concerns, this Part develops some general guidelines for randomized experimentation (without repeating all the advice developed in the previous section), describes how legislatures and administrative agencies might initiate randomization 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, there should be a large enough sample to generate meaningful results.¹⁵¹ 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 higher the number of observations, the better chance that any actual effect will correctly be identified as existing at any particular threshold of statistical significance. Policymakers need not, however, choose any particular level of statistical significance, such as 0.05, as a threshold for identifying an experiment as a success. Statisticians have long recognized these thresholds as arbitrary.¹⁵²

Meanwhile, policymakers must consider the unit of analysis at which randomization occurs.¹⁵³ 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

¹⁵⁰ J. King & R. Nicholson, *Informed Consent*, 3 INST. MED. ETHICS BULL. 1 (1986).

We might be more willing if we knew that the trial would increase our probability of getting the more effective therapy—but in this example, self-interested patients would prefer 100% of the therapy that is more likely to be effective.

¹⁵¹ See, e.g., DUFLO ET AL., *supra* note 87, at 29 (discussing the issue of sample size in randomized experiments).

¹⁵² See, e.g., Lester V. Manderscheid, *Significance Levels, 0.05, 0.01, or ?*, 47 J. FARM ECON. 1381 (1965) (urging that the applicable level of statistical significance be tailored to a particular purpose).

¹⁵³ See, e.g., DUFLO ET AL., *supra* note 87, at 40.

policies, but only at the risk of reintroducing omitted variable bias. Finally, policymakers should generally use matched samples, with matching occurring before the experiment on all available variables, to reduce attrition bias.

A final, but more controversial, design suggestion is to avoid problems of self-selection and attrition by making participation mandatory. Social experiments to date have largely been opt-in, allowing individuals to choose whether to participate and then perhaps also whether to opt out.¹⁵⁴ This is not surprising given the conventional view of social experimentation as a form of academic research. Academics cannot experiment on research subjects without informed consent.¹⁵⁵ But governments in theory could make participation in a randomized experiment mandatory (just as it has done with the draft lotteries) and even institute reporting requirements. There will always be some people who ignore the rules, and some unavoidable attrition, due to factors like emigration and death. But a government could either not count such individuals (and their matches) or develop some other convention for how to count them.¹⁵⁶

After accounting for experimental implementation costs, which are fixed, 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 experiments apply to 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 ordinary policy changes. When the policy environment does not change radically over time, this information 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

¹⁵⁴ Alice M. Rivlin & P. Michael Timpane, *Introduction and Summary*, in *ETHICAL AND LEGAL ISSUES OF SOCIAL EXPERIMENTATION*, *supra* note 97, at 1, 7.

¹⁵⁵ See generally Kathryn A. Tuthill, *Human Experimentation*, 18 J. LEGAL MED. 221 (1997) (providing a legal overview).

¹⁵⁶ The convention might depend on context. For example, in an experiment on securities disclosure, the bankruptcy of a corporation could count as stock price declining to zero. An individual’s death might count as a bad result in a health care policy experiment, but simply be ignored in an experiment on fee shifting in court cases. 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.”

from the pharmaceutical industry. In a drug trial, some individual subjects are given the experimental drug, while other individuals serving as controls receive the drug that constitutes the existing state of the art. 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.¹⁵⁷ 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 be theoretically 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 a different speed limit, with the limit applying to all individuals within the jurisdiction.

So how should policymakers decide 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. More fine-grained units of randomization are generally preferred so long as 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.¹⁵⁸ Thus, random assignment to individuals would be the best strategy when a

¹⁵⁷ RICHARD H. THALER & CASS R. SUNSTEIN, *NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS* 159-74 (2008).

¹⁵⁸ When policy is randomized at the state level, for example, serial correlation in error terms makes standard errors wide, and

policy targets individual outcomes and there are no spillovers to (untreated) other 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 in response to higher speed limits. 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 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,¹⁵⁹ then we should expect untreated firms that need not comply with the rule would be placed at a competitive advantage. 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. But when the outcome of interest concerns dimensions of social welfare that are not fully felt by the firms and their customers, the impact of the treatment on the untreated firm’s behavior may undermine analysts’ ability to parse out 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 time period in which to conduct the experiment. Longer experimental periods offer some obvious advantages. Long periods increase the chance that all involved parties become aware of the experiment and reduce the ability of the parties to avoid experimental effects by delaying

therefore complicates the finding of statistically significant policy impacts. For details, see Marianne Bertrand, Esther Duflo & Sendhil Mullainathan, *How Much Should We Trust Differences-in-Differences Estimates?*, 119 Q. J. ECON. 249 (2004).

¹⁵⁹ 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 market impact of efficient and inefficient mandates); *see also* Christine Jolls, *Accommodation Mandates*, 53 STAN. L. REV. 223 (2000).

behavior until the experiment completes. Both factors mean that longer periods are more likely to provide better estimates of the true effects of an experimental policy than short periods. At the same time, however, long-term experiments exacerbate the inequalities created by experimentation. In addition, experimental policies will often prove to be failures. 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.¹⁶⁰ Once it becomes clear that one treatment is preferred to another, it is immoral and inefficient to capriciously expose subjects to the inferior policy. In other circumstances it will be appropriate to extend the length of the experiment to gather more information. This is especially true with regard to multi-level randomized testing, where 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 provide a vivid example of this form of “convexification” with regard to Internet ads. If a randomized experiment initially suggests that “Tastes Great” is a more successful beer ad than “Less Filling,” the Google software will automatically start increasing the probability that people will see the more successful ad. This method—which is called “outcome-adaptive randomization”—mitigates the inefficiency of additional testing, but allows the researcher to continue to collect some information on the longer-term effects of the various policy treatments.¹⁶¹

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

¹⁶⁰ For example, the National Institute 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, February 23, 2007, at A3.

¹⁶¹ Ying Kuen Cheng, Lurdes Y.T. Inoue, J. Kyle Wathen, Peter F. Thall, *Continuous Bayesian Adaptive Randomization Based on Event Times with Covariates*, 25 STAT. IN MED. 55 (2006).

randomization becomes more common, an executive order might insist that agencies systematically consider what policies should be randomized. Meanwhile, in legislatures, there is a danger that even solid evidence produced by randomized experiments will be ignored, and this 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 to simply ignore experimental results.

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 agencies themselves could decide to randomize policies. The courts would presumably examine such a decision with the usual tools of administrative judicial review, ensuring for example that the action was procedurally proper,¹⁶² was consistent with law,¹⁶³ and represented a permissible policy judgment.¹⁶⁴

These hurdles should be straightforward for an agency to clear. As long as an agency goes through the ordinary notice-and-comment process,¹⁶⁵ providing a detailed explanation of the purpose of an experiment in the notice of proposed rulemaking,¹⁶⁶ as well as a “concise, general statement” of basis and purpose,¹⁶⁷ 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 nor the control legal regimes is inconsistent with the agency’s governing statute, a decision to launch an experiment should present no problem for *Chevron* review. Perhaps the most significant obstacle would be hard look review,¹⁶⁸ in which a court would examine the

¹⁶² See, e.g., 5 U.S.C. § 553 (2006) (setting forth requirements for notice-and-comment rulemaking).

¹⁶³ See, e.g., *Chevron U.S.A., Inc. v. Natural Resources Defense Council, Inc.*, 467 U.S. 837 (1984) (setting forth the modern standard for evaluating agency legal interpretations).

¹⁶⁴ See, e.g., 5 U.S.C. § 706(2)(A) (2006) (requiring the reviewing court to “hold unlawful and set aside agency action . . . found to be . . . arbitrary, capricious, an abuse of discretion, or otherwise not in accordance with law”).

¹⁶⁵ *Id.* § 553(b) (requiring publication of general notice of proposed rulemaking); *id.* § 553(c) (requiring opportunity for comment and issuance of concise general statement of basis and purpose).

¹⁶⁶ Agencies generally seek to meet the “general notice” requirement by publishing the actual rules that they are considering implemented, though even this is sometimes inadequate. See, e.g., *Portland Cement Ass’n v. Ruckelshaus*, 486 F.2d 375 (D.C. Cir. 1973) (vacating a rulemaking for failure to release background documents necessary to be able to respond to notice).

¹⁶⁷ “Concise” and “general” are sometimes interpreted to meet “detailed” and “specific.” See *Automotive Parts & Accessories Ass’n v. Boyd*, 407 F.2d 330, 338 (D.C. Cir. 1968) (warning against interpreting these words literally).

¹⁶⁸ See, e.g., *Greater Boston Television Corp. v. FCC*, 444 F.2d 841, 851 (D.C. Cir. 1970) (“Its supervisory function calls on the court to intervene . . . if the court becomes aware . . . that the agency has not really taken a ‘hard look’ at the salient problems, and has not genuinely engaged in reasoned decisionmaking.”).

agency's justification for creating the experiment. But hard look review is supposed to be deferential,¹⁶⁹ and an agency should be able to justify employing a randomized experiment on the ground that this approach could provide information relevant to the administrative process.

Indeed, an administrative agency should perhaps receive broader latitude to create an experiment than to create a new administrative regime without an experiment. Procedurally, an agency might argue that it should not have to go through notice-and-comment to establish an experiment,¹⁷⁰ 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,¹⁷¹ perhaps in part because this would allow an administrative agency to renew a program indefinitely.¹⁷² An agency should, however, at least be allowed to focus solely on the reason for conducting an experiment, rather than responding to comments on the merits of the underlying policy issue. Because an experiment produces data on a policy issue, courts should not require an agency to show that existing data already justifies the policy that the experiment is designed to test.

To enact an experimental policy permanently, an agency presumably would face hard look review, but here too the courts should perhaps be more deferential than usual. Critics of notice-and-comment have complained that it is "ossified,"¹⁷³ 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.¹⁷⁴ An agency conducting an

¹⁶⁹ See, e.g., *Motor Vehicle Manufacturers Ass'n v. State Farm Mutual Automobile Ins. Co.*, 463 U.S. 29, 43 (1983) ("The scope of review under the 'arbitrary and capricious' standard is narrow and a court is not to substitute its judgment for that of the agency.").

¹⁷⁰ This might be so when Congress has explicitly instructed an agency to conduct an experiment. Even here, however, the agency is generally like to notify the public of its intent to run an experiment, for example by soliciting third parties to perform the experiments. See, e.g., *Medicare Program; Solicitation for Proposals for the Demonstration Project for Disease Management for Severely Chronically Ill Medicare Beneficiaries with Congestive Heart Failure, Diabetes, and Coronary Heart Disease*, 67 FED. REG. 8267-001, 2002 WL 245888.

¹⁷¹ The Administrative Procedure Act continues a general exemption for notice-and-comment "when the agency for good cause finds . . . that notice and public procedure thereon are impracticable, unnecessary, or contrary to the public interest"). 5 U.S.C. § 553(b)(B). But the courts have found that the temporary nature of a rule is not enough to escape notice-and-comment. See *Tenn. Gas Pipeline v. FERC*, 969 F.2d 1141 (D.C. Cir. 1992).

¹⁷² But see Juan J. Lavilla, *The Good Cause Exemption to Notice and Comment Rulemaking Requirements Under the Administrative Procedure Act*, 3 ADMIN. L.J. 317, 378 (1989) (suggesting that courts enforce temporary rules only until the agency has had enough time to develop a permanent rule, whether or not it has done so).

¹⁷³ See, e.g., Thomas O. McGarity, *Some Thoughts on "Deossifying" the Rulemaking Process*, 41 DUKE L.J. 1385 (1992); Richard J. Pierce, Jr., *Seven Ways to Deossify Agency Rulemaking*, 47 ADMIN. L. REV. 59 (1995).

¹⁷⁴ An ideological agency, facing an ideologically hostile court, might respond either by investing more in meeting the requirements of the hard look doctrine or "allocate their resources to other projects." Richard L. Revesz, *Environmental*

experiment is less likely to be following an ideological agenda than an agency drawing inferences based on existing data that plausibly might support different conclusions. Moreover, courts conducting judicial review should recognize the unique value of evidence from randomized experimentation.¹⁷⁵ There remains a danger that an agency might make invalid inferences on the basis of an experiment. At least where experiments provide the best available evidence on a policy issue, however, courts should allow an agency to reply that it placed more weight on the experimental evidence, without chronicling on a case-by-case basis all of the problems of nonexperimental evidence.

If individual agency initiatives begin to test and choose policies with experimental means, randomization could gradually become a more entrenched part of the policy process. Perhaps someday, considering randomization might become almost as routine and formalized as cost-benefit analysis.¹⁷⁶ For example, the executive branch might provide a standard procedure for agencies to consider randomization and to produce a randomization impact statement (RIS) when enacting a new rule, whether or not the agency decided to use a randomized approach.

An RIS might include the following elements:

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 minimus 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 quantitatively measure 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)¹⁷⁷ to resolve the legislative debate, because legislators and their constituent groups may have incommensurable preferences.¹⁷⁸

Regulation, Ideology, and the D.C. Circuit, 83 VA. L. REV. 1717, 1770 (1997).

¹⁷⁵ Cf. Dorf & Sabel, *supra* note 2, at 397 (arguing that hard look review should reward experimental agency approaches, though not focusing specifically on random experimentation).

¹⁷⁶ See generally Executive Order No. 12866 § 1(b)(6) (1993) (permitting regulation only where benefits exceed costs).

¹⁷⁷ See *Planned Parenthood of S.E. Pa. v. Casey*, 505 U.S. 833 (1992) (striking down Pennsylvania spousal notification law).

RANDOMIZING LAW

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 just the agency's examination of the data generated by the experiment, 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 rather than 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 randomization impact statements. Once again, no doctrinal innovation is necessary here, as the courts could assess randomization impact statements with the usual tools of hard look review, ensuring that agencies have carefully addressed counterarguments both to decisions whether to engage in randomization and in decisions after experimentation occurs. Creation of the randomization impact statement 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 randomization impact statement could also provide an opportunity for the Office of Management and Budget or some other specialized agency to generate expertise in administering experiments that could then be applied to experiments from all agencies. Such an 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

But a state might experiment on offering the possibility of giving 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 (2008).

¹⁷⁸ Then again, some moderate legislators might be swayed by compelling evidence about the impact of notification law on: i) a women's propensity to abort; ii) the propensity of unaborting fetuses to commit crime; and iii) the probable psychological well-being of the spouses. See Cass R. Sunstein & Adrian Vermeule, *Is Capital Punishment Morally Required? Acts, Omissions, and Life-Life Tradeoffs*, 58 *STAN. L. REV.* 703 (2006); 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.

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.¹⁷⁹

2. *Legislatures: The Case for Self Execution*

Given the problems identified in Part IV, random experimentation data will rarely give unambiguous answers to multi-dimensional 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.¹⁸⁰ 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 so it cannot avoid these obstacles altogether. But if through gradual steps randomized self-executing experiments become sufficiently familiar that they no longer seem strange, then a culture of random legal experimentation might slowly emerge. Legal experiments should be easier to enact in this culture than are legal reforms in our present legal culture. A marginal decisionmaker, uncertain whether to support a program, should be more willing to favor it when the program will continue if and only if it turns out to be successful. Supporters of a program, meanwhile, may find it difficult to oppose a measure that would condition continuation of the program on confirmation of its success. Even opponents of moving the law in the direction of an

¹⁷⁹ One prominent clinical trial company has run over “3,200 trials in some 50 countries” since the year 2000. *See* Working with Quintiles, <http://www.quintiles.com/AboutUs/WorkingWithQ.htm> (last visited Sept. 5, 2008).

¹⁸⁰ *See infra* Part III.A.1.

experiment might nonetheless be willing to support the experiment if they believe that it will turn out to be a failure.

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

The principal challenge of self-execution is determination of what counts as success. The metrics they agree on may well be simple proxies, far less sophisticated than what statisticians *ex post* would rely on. 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 of all 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.

This highlights that we cannot consider legal experiments solely as a social scientist might. Rather, we must consider legal experimentation as a mechanism of the policymaking process, an imperfect device for converting scientific knowledge into law. Sometimes, the criteria of scientific usefulness and legal practicality point in different directions. An experiment might be beneficial even if its results add little to social science knowledge; a simple randomization scheme may be beneficial even where econometricians would prefer a more elaborate treatment design; and an experiment might compare two legal approaches varying along a number of dimensions even though this may make the results difficult to interpret. 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, that decisionmaker might be biased, for example in favor of the executive's policy preferences.

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.¹⁸¹ Moreover, 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.¹⁸² 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 long-standing theoretical debates.

a) *A Short Sale Experiment*

Policymakers recently have begun to grasp the potential of randomized policy experiments for securities. In 2004, the SEC issued Rule 202T of Regulation SHO, devising an experiment to test some restrictions on short sales.¹⁸³ Scholars have debated the effect of these restrictions. Finance theory predicts that short sale restrictions should reduce the volume of short selling, which in turn should reduce the liquidity of a stock and potentially lead to less accurate pricing.¹⁸⁴ Others argue that the restrictions help prevent stock manipulation by coordinated short sellers seeking to force the price of a stock down simply to purchase it a low price.

Rule 202T allowed the SEC to implement a “pilot program to examine the efficacy” of the short sale restrictions.¹⁸⁵ The pilot program exempted one third of the stocks in the Russell 3000 from the short sale restrictions. The exempted stocks were chosen by sorting the 2004 Russell 3000 first by listing market [e.g., NYSE, NASDAQ], then by average daily dollar volume from June 2003 through May 2004, and then selecting every third company starting with the second. This is an example of stratified sampling.¹⁸⁶ So long as it is effectively random which

¹⁸¹ The lack of interstate variation explains the intense empirical interest in the relatively infrequent change in securities laws at the national level. For example, the 1930s, when modern securities law was first introduced, continue to be an active area of research, as is an expansion of the securities law regime to over-the-counter (OTC) stocks in the 1960s. Michael Greenstone, Paul Oyer & Annette Vissing-Jorgensen, *Mandated Disclosure, Stock Returns, and the 1964 Securities Acts Amendments*, 121 Q.J. ECON. 399 (2006); Paul G. Mahoney, *The Political Economy of the Securities Act of 1933*, 30 J. LEGAL STUD. 1 (2001); Allen Ferrell, *Mandated Disclosure and Stock Returns: Evidence from the Over-the-Counter Market* (Harvard Law & Economics Discussion Paper No. 453, 2003).

¹⁸² Stephen E. Christophe, Michael G. Ferri & James J. Angel, *Short-Selling Prior to Earnings Announcements*, 59 J. FIN. 1845 (2004); Ian Ramsay, *Short-Selling: Further Issues*, 21 SEC. REG. L.J. 214 (1993).

¹⁸³ OFFICE OF ECON. ANALYSIS, U.S. SEC. AND EXCH. COMM’N, ECONOMIC ANALYSIS OF THE SHORT SALE PRICE RESTRICTIONS UNDER THE REGULATION SHO PILOT 3 (2007), available at <http://www.sec.gov/news/studies/2007/regshopilot020607.pdf> (hereinafter “SEC Report”) (describing the restrictions).

¹⁸⁴ *Id.* at 6-8.

¹⁸⁵ *Id.* at 4.

¹⁸⁶ Stratified sampling occurs because:
in any randomized trial it is desirable that the comparison groups should be as similar as possible as regards participant

of 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 this experimental design 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 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.¹⁸⁷ The Office of Economic Analysis of the SEC produced a comprehensive report on the pilot program, with many of the components that we recommend for the RIS. The report first reviews the theoretical and empirical literature on short sale restrictions. This literature tends to view the existing policy of short sale restrictions as inefficient. 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.”¹⁸⁸ 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.¹⁸⁹

The report then provides a detailed explanation of how the experiment was conducted, with a discussion and justification of the stratified sampling method used in the experiment.¹⁹⁰ In addition, the report explains the methodological tools applied to examine the impact of the short sales restrictions on various outcomes.¹⁹¹ Finally, the report provides a detailed examination of

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.

Stratified Randomization, Evidence-Based Medicine, Glossary of Terms, <http://www.sahealthinfo.org/evidence/s.htm> (last visited Sept. 5, 2008). 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 volume.

¹⁸⁷ *SEC Report, supra* note 183, at 4.

¹⁸⁸ *Id.* at 4.

¹⁸⁹ *Id.* at 6-10.

¹⁹⁰ *Id.* at 22-28.

¹⁹¹ *Id.* at 28-34.

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.¹⁹² The report examines each outcome variable of interest, and finds that eliminating short sale restrictions impacts some outcome variables (such as short selling volumes, which are approximately 8% 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).¹⁹³ 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.¹⁹⁴ The report concludes that:

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

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 a third of the firms in the Russell 3000 highlights the value of experiments for policymaking. The experiment demonstrated that the 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 restriction reduces freedom without 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. There is no

¹⁹² *Id.* at 34-51.

¹⁹³ *Id.* at 51-56; *see also id.* tbls. 3 & 6.

¹⁹⁴ *See id.* Appendix A.

systematic effect of the restriction 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 section proposes such an experiment in one area—the Sarbanes-Oxley law, but experiments can apply to any controversial issue.

b) Experimental Sarbanes-Oxley Repeal

In the wake of the Enron/WorldCom accounting scandals in 2002, Congress passed the Sarbanes-Oxley Act (SOX). SOX included many provisions to improve the quality of financial reporting and corporate governance. Some of SOX’s prominent provisions include mandatory CEO and CFO certification of financial results and new “internal controls” requirements.¹⁹⁵

SOX has proven quite controversial. Many corporations and academics dispute SOX’s efficacy in preventing fraud, while bemoaning its expense. Others argue that SOX performs a critical role in improving confidence in financial markets. This debate has spawned an extensive empirical literature evaluating SOX’s impacts on corporate value, cross listing in the U.S. markets, and going-private decisions.¹⁹⁶ Many empirical papers find that SOX appears to destroy value or reduce cross listings, but these findings are disputed by others.

The ambiguity about SOX’s desirability is reflected in calls for SOX’s elimination. To this point, however, SOX’s proponents have managed to prevent any alteration of SOX. SOX offers an almost ideal context for a randomized repeal of securities legislation. SOX’s provisions may well destroy value, 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.

¹⁹⁵ The internal controls requirements obligated companies to set up elaborate mechanisms for detecting malfeasance within the company or disclose the absence of such controls.

¹⁹⁶ Peter Iliev, *The Effect of SOX Section 404: Costs, Earnings Quality and Stock Prices*, 65 J. FIN. 1163(2010); see also Ellen Engel, Rachel M. Hayes & Xue Wang, *The Sarbanes-Oxley Act and Firms’ Going-Private Decisions*, 44 J. ACCT. & ECON. 116 (2007); Roberta Romano, *The Sarbanes-Oxley Act and the Making of Quack Corporate Governance*, 114 YALE L.J. 1521 (2005); Ivy Xiying Zhang, *Economic Consequences of Sarbanes-Oxley Act of 2002*, 44 J. ACCT. & ECON. 74 (2007).

markets may have declined because of SOX's onerous requirements, or they may have declined due to the development of foreign exchange's sophistication, decreasing the value of U.S. markets as a source of capital. An experimental repeal of SOX for some companies is likely to provide convincing empirical evidence that resolves which of these factors is more important. Moreover, because SOX is so unpopular with corporations, instituting an experimental repeal should prove popular, while avoiding the political battle that would be caused by attempting to permanently repeal SOX for all companies.

Randomized experimental repeal of SOX should take place as follows. First, the most controversial provisions of SOX 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 SOX restrictions repealed) and control groups (with SOX continuing as presently). For example, foreign companies cross listed in U.S. markets should be well represented in both the sample and control groups to help evaluate SOX's effect on delisting from U.S. markets. The experimental repeal period should be a relatively long one. Many of SOX's effects will only be felt 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 SOX repeal will not provide a good test of SOX's true effects.¹⁹⁷ Instead, the experimental repeal should be applied for an extended period—up to several years.¹⁹⁸

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

¹⁹⁷ Because market values incorporate expectations of future profits, market values respond very quickly to the impact of 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 SOX 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 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.

¹⁹⁸ While this might appear to be a long period, the status quo, with a controversial law applied indefinitely, is in many ways just as speculative as an experiment, but without producing information that would yield policy conclusions.

publicly traded corporations, making the choice of unit of observation relatively straightforward. If SOX repeal is likely to produce substantial competitive advantages for untreated firms (i.e., those still subject to SOX requirements),¹⁹⁹ then the unit of randomization may need to be raised to the industry level. 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 SOX rather than on SOX as a whole. Thus, some companies would have the internal control provisions eliminated, but other provisions of SOX 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 SOX intact, and so on. Randomizing different permutations of the controversial provisions in SOX allows for the identification of specific provisions that are effective or ineffective, rather than 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.²⁰⁰

Because many companies find SOX compliance costly and are likely to volunteer, a test of SOX could ask for companies to volunteer to participate in a SOX repeal experiment and then assign some of these companies to a treatment SOX-repeal group and others to a control group with SOX remaining in place.²⁰¹ The experiment with volunteer companies would provide a good estimate of the treatment effect of allowing companies to opt out of SOX, because companies that volunteer to take part in an experimental repeal of SOX are likely to be similar to

¹⁹⁹ For example, suppose that investors benefit from the improvement in information quality mandated by SOX, but that investors can apply this information from companies subject to SOX to companies not subject to SOX. In this case, the non-SOX companies may do better than the SOX companies because they get the benefit of the improved information without incurring its expense. This difference in outcomes, however, does not provide an accurate estimate of the impacts of a full SOX repeal. If no companies followed SOX, then there would be no informational spillovers and all companies might be worse off. An experiment which 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.

²⁰⁰ An interaction effect occurs when the effect of one variable is dependent upon the value of another variable. For example, CEO certification provisions taken alone might have no impact on 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 so that the combination of the two provisions has an impact on value.

²⁰¹ Repealing SOX for all companies that volunteer for the SOX repeal experiment and estimating the impact of SOX by comparing these companies with companies that did not volunteer for the experiment (for whom SOX remained in place) fails to provide accurate estimates of the impact of SOX. Companies that volunteer for SOX 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 SOX. As a result, some companies that volunteer for SOX repeal should be randomly assigned to a control group that must remain SOX compliant. These companies will be similar to the companies that volunteered for a SOX repeal and had SOX repealed, making estimates of the effect of a SOX repeal more accurate.

RANDOMIZING LAW

companies that would opt out of SOX, were that an option. Examining an experiment with volunteers would provide a poor estimate of the impact of a full repeal of SOX, however, because the impact of SOX on companies that volunteer to have SOX eliminated is likely to be different from the impact of SOX on the average company.²⁰²

To estimate the impact of a full SOX repeal on the average company, SOX repeal could be randomly but mandatorily assigned to some companies but not others. This would incur the cost of forcing some companies to experience SOX repeal unwillingly, but avoids the problem of estimating the impact of SOX exclusively for companies that volunteer to have SOX repealed. A randomized mandatory repeal of SOX 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.

There are many potential outcomes of interest for a SOX-randomized experiment. SOX aimed to restore investor confidence in the financial markets and financial reporting. One obvious outcome variable is therefore investor confidence in the quality of corporate reporting. A related measure would include the amount of fraud in SOX companies relative to non-SOX companies. To financial economists, however, confidence and prevention of fraud are not aims but rather means to an end. 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 SOX compliance. Therefore, other measures that account for both the costs and benefits of SOX should be examined.

One important alternative measure of SOX's efficacy is stock market value. Stock market value goes up if investors perceive that SOX reduces the cost of capital and costs nothing, but goes down if SOX 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 SOX's net effects. Because of the randomized nature of a SOX experiment and the large number of companies that would participate, a long term study of the

²⁰² See *supra* note 119.

impact of SOX on market value is possible, providing evidence not just of the market's impressions of SOX, but also of the market's verdict after observing SOX's impacts. If, after a number of years, SOX companies have outperformed non-SOX companies, then this constitutes solid evidence that SOX enhances corporate value.

If a SOX 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 particular features of the SOX repeal should be retained. The case for self-execution is particularly strong if it appears likely that Congress otherwise might well ignore the experiment, with partisans sticking to their original positions regardless of the experimental results. Even a perfect experiment will not resolve all questions about SOX t, for example because partisans of one position or another might argue reasonably that the result could have been different if the experiment 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. An imperfect proxy may be more likely to produce beneficial legislative change than 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 SOX, the SEC can 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 suggested here for SOX, 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.²⁰³ Others argue

²⁰³ 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

that hours and entrepreneurship are not particularly sensitive to relatively small changes in marginal tax rates, meaning that government revenues will fall nearly proportionately to the amount of a tax decrease. These arguments are rehashed whenever the government considers raising or lowering taxes (in other words, almost annually).²⁰⁴

Tax rates change frequently, so there is ample variation with which to study how the change in tax rates impacts labor supply and entrepreneurship.²⁰⁵ Unfortunately, these changes in rates are often correlated with many other things, making it extremely difficult to draw firm conclusions about the response of labor supply to tax rates. For example, tax rates are often altered in response to changes in economic conditions.²⁰⁶ If economic behavior changes after rates change, the changes may be attributable to the change in rates, or it may be attributable to the changing economic conditions that motivated the change in rates. Such confounding factors help explain the lack of consensus about the true impact of taxes on labor supply incentives.²⁰⁷

Randomized experimental manipulation of tax rates will not suffer from this complication. If tax rates are randomized at the individual level, then individuals facing very similar economic conditions will be subject to different tax rates. If these individuals behave differently, then the 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, but subject to different marginal tax rates. If the individual subject to the lower tax rates works many more hours than her counterpart subject to higher tax rates, then this 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.²⁰⁸ The critical outcomes of interest in the tax debate is the impact of tax rates on labor supply and

economists even claim that lowering tax rates can increase revenue, but this claim is discredited. N. Gregory Mankiw, *The Optimal Collection of Seigniorage Theory and Evidence*, 20 J. MONETARY ECON. 327 (2004).

²⁰⁴ David Rosenbaum, *Economic View: Name That Tune About Tax Cuts*, N.Y. TIMES, May 18, 2003, at 3; Glenn Kessler, *Now President Faces Tax Cut Test; Loss of Revenue Means Bush Needs to Cut Spending*, WASH. POST, Feb. 11, 2001, at A5.

²⁰⁵ See, e.g., DANIEL J. MITCHELL, THE HERITAGE FOUNDATION, LOWERING MARGINAL TAX RATES: THE KEY TO PRO-GROWTH TAX RELIEF (2001), available at <http://www.heritage.org/research/taxes/BG1443.cfm>; Basil Dalamagas, *The Effects of Tax Rate Changes on Output and Government Deficits*, 10 APPLIED ECON. LETTERS 97 (2003).

²⁰⁶ See, e.g., David M. Herszenhorn, *Bush and House in Accord for \$150 Billion Stimulus*, N.Y. TIMES, Jan. 25, 2008, at A1 (describing 2008 tax rebate).

²⁰⁷ Again, this 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 (the Laffer curve), for example, would be rejected by the vast majority of serious scholars.

²⁰⁸ Note that by varying the unit of randomization between the individual and the household, policymakers can get a sense of the

entrepreneurship. These decisions are made at the individual or household level, meaning that individuals or households are the appropriate units of observation.²⁰⁹

Imposing differential mandatory tax rates on similarly situated individuals might be controversial. One response 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.²¹⁰ 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, that 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 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 impacts of differential marginal tax rates *and* offsetting transfers. If transfers also have an effect on labor supply—such as a wealth effect²¹¹—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. If tax rates change for a brief time, individuals subject to low tax rates may shift work from future periods into the

true effect of the “marriage penalty,” James Alm, Stacy Dickert-Conlin & Leslie A. Whittington, *Policy Watch: The Marriage Penalty*, 13 J. ECON. PERSP. 193 (1999), and other important questions of tax policy.

²⁰⁹ If policymakers want to study the spillover effects of taxes, such as whether 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 startup businesses founded by those with lower tax rates and estimate the identity and salaries of employees of the startup business. If this proves impossible, then tax rates can be randomized at other units of observation, such as the state or county.

²¹⁰ Credit for the idea of a state-sponsored lottery for individuals who meet tax law requirements belongs to Terrence Chorvat.

²¹¹ Alan B. Krueger & Jorn-Steffen Pishke, *The Effect of Social Security on Labor Supply: A Cohort Analysis of the Notch Generation*, 10 J. LABOR ECON. 412 (1992).

current period in order to take advantage of the lower tax rate. If people do this, the experiment will generate an unrealistically high estimate of the impact of tax rates on labor supply; the experiment will reflect abilities to shift work between time periods rather than to permanently adopt different labor arrangements in response to different incentives. A longer experimental period limits the ability of individuals to shift work between periods. Work can be moved 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 (e.g., 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 variable is labor supply and wages. The experiment will directly address the degree to which lower taxes induce individuals to work more hours or seek more demanding higher wage jobs. Many other outcome variables, such as entrepreneurship levels, child 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 may be particularly attractive if Democrats and Republicans on average genuinely 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 partisans on the losing side can be expected to conjure some explanation for losing instead of changing their views on taxes.

3. Civil Rights

Up to this point, most of our examples have concerned experiments concerning corporate or public finance. But the idea of randomized testing could be applied to a much larger set of laws that more directly concern the regulation of individual behavior. This section 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

RANDOMIZING LAW

orientation.²¹² The Employment Non-discrimination Act (ENDA)—a minimalist prohibition of disparate treatment on the basis of sexual orientation—has been introduced in Congress several times, recently passing in the House in 2007.²¹³ Even though polls suggest that an overwhelming majority of Americans oppose employment discrimination on the basis of sexual orientation,²¹⁴ opponents argue that ENDA would impose substantial litigation and other compliance costs that would be visited on private employers.²¹⁵

A 2000 GAO study sheds some light on the question of litigation costs by analyzing the number of claims that had been made in the eleven states that had prohibited sexual orientation discrimination by private employers as a matter of state law.²¹⁶ One of us analyzed the claims data together with more general employment data and found that historically each year there has only been about one claim for every 60,000 workers.²¹⁷ If the employer's average cost per complaint were \$100,000, the average annual cost of the statute per employee would be less than \$2.²¹⁸

While this analysis of historic data suggests that employer costs are quite low, there is a chance that these estimates might not represent the costs that a federal law would produce. For example, it is possible that employers in the first eleven states to pass the law are less likely to discriminate than those in the remaining thirty-nine. Or it might be possible that the specific language of ENDA would produce lower (or higher) costs of compliance than the state statutes. A randomized test of the impact of ENDA is a natural and powerful way to learn more about

²¹² Twenty states and the District of Columbia have passed state statutes that prohibit employers from discriminating on the basis of sexual orientation. HUMAN RIGHTS CAMPAIGN, STATEWIDE EMPLOYMENT LAWS AND POLICIES (2008), *available at* http://www.hrc.org/documents/employment_laws_and_policies.pdf.

²¹³ *See, e.g.*, H.R. 3017, 111th Cong. (1st Sess. 2009).

²¹⁴ Gallup Poll, Question Id: USGALLUP .0331 Q19; *see also* 2004 L.A. Times Poll, Question Id: USLAT .041104 R52 (70 percent favor . . . laws to protect gays against job discrimination); GLAAD Media Reference Guide, <http://www.glaad.org/media/guide/infocus/polls.php> (last visited Sept. 6, 2008) (Gallup poll in 2005 shows 87% support); John Newsome, On Protecting Gay Americans from Workplace Discrimination Employment Non-Discrimination Act (ENDA) Vote Tests Our Values, S.F. CHRON., Nov. 7, 2007, *available at* <http://www.sfgate.com/cgi-bin/article.cgi?file=/c/a/2007/11/07/EDD7T6UMC.DTL> (2006 Gallup poll shows 89% support).

²¹⁵ *See, e.g.*, Statement of Administration Policy, H.R. 3685 (Oct. 23, 2007), *available at* <http://www.whitehouse.gov/omb/legislative/sap/110-1/hr3685sap-r.pdf> (policy position of former President George W. Bush).

²¹⁶ Letter from the GAO to the Honorable James M. Jeffords, Chairman, Committee on Health, Education, Labor and Pensions, Sexual Orientation Based Employment Discrimination: States' Experience with Statutory Prohibitions Since 1997 (Apr. 28, 2000).

²¹⁷ Ian Ayres & Jennifer Gerarda Brown, *Mark(et)ing NonDiscrimination: Privatizing ENDA with a Certification Mark*, 104 MICH. L. REV. 1639, 1645 (2006) (tbl. 1: Analysis of Litigation Rates and Expected Costs of State Prohibitions).

²¹⁸ *Id.*

RANDOMIZING LAW

whether the 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 if ENDA caused covered firms to lose market share to uncovered firms.

In this subsection, we discuss how such a test might be structured. Although it would be theoretically possible to randomly assign the application of ENDA 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. So 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. (By comparing relative market shares of the covered and uncovered firms, analysts can test for any impact on competition.) Firms assigned to the status quo control group (no prohibition of discrimination) might, however, be impacted by the treatment group, if employees transferred to or from the treatment group because of the discrimination prohibition. If concern over this type of overflow effect is large enough, it might militate for randomizing at the industry level—or conducting a mixed experiment that partially randomizes at the industry and partially at the firm level.²¹⁹

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 seven million businesses with employees²²⁰—that it would be possible to perform a powerful test that assigns perhaps 1% 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,

²¹⁹ Alternatively, the possible overflow effects of employees could be dampened by randomizing across cities or states. But the plausible size of this impact to our minds would not justify reducing the number of observations.

²²⁰ U.S. CENSUS BUREAU, CENSUS, tbl. 735, *available at* <http://www.census.gov/compendia/statab/tables/08s0735.pdf>.

RANDOMIZING LAW

as they would have if ENDA passed. Indeed, Jennifer Brown and Ian Ayres have created a contractual mechanism where any employer with just a few clicks at www.fairemploymentmark.org can do just that.²²¹ 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 in to 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.²²²

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.²²³ Instead of running an experiment on the effects of mandatory ENDA, it would be possible to test the impact of varying the default or menu dimensions of the law. Specifically, we could imagine randomizing firms into three groups: a control group with the status quo federal coverage; an "opt in" group of firms that could affirmatively opt for coverage by sending a notice to the Justice Department; and an "opt out" group of firms that could avoid liability under the statute by sending notice (in advance of any claimed discrimination) to the Justice Department that they did not wish to be covered.²²⁴

VI. CONCLUSION

Randomized experimentation offers a powerful means of evaluating 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

²²¹ The fair employment license falls short of ENDA protections in a few dimensions. See Ayres & Brown, *supra* note 217, at 23 (noting that the license would not be enforced by governmental agencies and private suits could not be brought in federal court); Ian Ayres & Jennifer Gerarda Brown, *Privatizing Employment Protection*, 49 ARIZ. L. REV. 587 (2007).

²²² Ian Ayres, *Menus Matter*, 73 U. CHI. L. REV. 3 (2006).

²²³ Yair Listokin, *What Do Corporate Default Rules and Menus Do? An Empirical Examination* (Yale Law School, Working Paper No. 335, 2005).

²²⁴ Randomized tests of default rules and menu options do pose particular problems for maintaining an uncontaminated control group similar to those described above. It is possible that the control group's behavior will be impacted by the treatment. Control group firms may be confused about the legal regime under which they are operating. 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. James Heckman, Neil Hohnmann, Jeffrey Smith & Michael Khoo, *Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment*, 115 Q. J. ECON. 651 (2000).

RANDOMIZING LAW

policies. The vast majority of policy changes, however, are enacted without the benefit of randomized evaluations. This Article seeks to systematize and expand the use of randomized experiments of law and policy. In the short term, a number of individual experiments could advance the cause 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 could encourage legislators to back up their empirical claims with a willingness to initiate experiments through legislation, with policy outcomes dependent on experimental results.