

# Randomizing Law

Ian Ayres\* and Yair Listokin\*\*

Government should embrace randomized trials of different laws and regulations as a tool for testing what works. Just as a 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 propose the presumptive use of “Randomized Impact Statements” as prerequisites for regulations and even statutes to systematically assess the likely impact of legal change. Randomized control trials of statutory change can be used to assess enactments of new statutes as well as proposed repeals of old statutes. We consider circumstances in which the randomization technique is not well suited to assess legal impacts, as well as specific examples where the technique could provide legislators with valuable information about whether the proposed change is warranted.

23,367

\*William K. Townsend Professor, Yale Law School, [ian.ayres@yale.edu](mailto:ian.ayres@yale.edu).

\*\*Associate Professor, Yale Law School, [yair.listokin@yale.edu](mailto:yair.listokin@yale.edu). Anthony Vitarelli provided excellent research assistance.

## Table of Contents

I.	Introduction.....	3
II.	The Power of Randomization .....	6
III.	A Proposal of Presumptive Randomized Testing of Legal Change .....	8
A.	Choosing Policies To Test .....	9
1.	Costs of Policy Experimentation .....	9
2.	Choosing Experimental Policy to Minimize Experimental Costs .....	10
3.	Weighing the Costs and Benefits of Experimental Policies .....	12
4.	Identification of Policies for Experimentation.....	12
B.	Experimenting with Policies.....	14
1.	Choosing the Unit of Randomization and Experimentation .....	14
2.	Choosing the Experimental Subjects .....	16
3.	Experimental Length.....	18
C.	Enacting Permanent Policies.....	19
IV.	Caveats.....	20
A.	Randomized Experiments that Are Not Double-Blind .....	21
B.	Difficulties of Estimating Policy Relevant Treatment Effects from Randomized Experiments .....	23
C.	The Bias of Incomplete Knowledge .....	24
D.	The (Im)Morality and (Un)Constitutionality of Randomized Experiments .....	25
E.	The Political Economy of Randomized Testing .....	29
V.	Applications .....	31
A.	Securities Law.....	31
1.	A Short Sale Experiment .....	31
2.	Experimental Sarbanes-Oxley Repeal .....	34
B.	Tax Law Experiments .....	38
C.	An ENDA Experiment.....	40
VI.	Conclusion .....	44

## 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.<sup>1</sup> These debates were theoretical in nature—with scholars predicting the positive and negative effects of monopolies versus a competitive market. Over two thousand years later, theoretical debates over policies as disparate as employment discrimination, securities regulation, and tax policy remain the norm. Many of these debates are long-lasting. For example, scholars have been debating the appropriate degree of disclosure regarding publicly traded securities for over fifty years. But many important policy issues cannot be resolved by theory alone, for the simple reason that equally good theories point in different directions. Do concealed handgun laws increase or decrease crime? Does affirmative action in legal education increase or decrease the number of minority attorneys? Theory can't resolve these important questions, because you can tell coherent theoretical stories to support different answers.<sup>2</sup>

The absence of conclusive evidence regarding the causal impacts of different policies hinders the resolution of these policy debates. Scholars attempt to inform these debates by trying to parse from historical data the impacts of differences in policy, but regression analysis of policy is fraught with complications. First, there is little policy variation on many topics of national importance. Securities laws, for example, have changed only incrementally since the 1930s. Second, the variation that does exist is correlated with many other factors. States that restrict guns, for example, may be different along many dimensions from states that have relatively permissive gun laws. Attributing differences in outcomes (such as crime rates) to differences in gun policy is therefore extremely difficult. While the outcome differences may be due to differences in gun policy among states with disparate gun policies, or they may also be caused by other differences between states. Empirical policy evaluation resembles a drug study where the experimental population gets to choose whether to take the medicine or the placebo.<sup>3</sup>

In this Article, we argue that government should embrace randomized trials of laws and regulations as a tool for testing what works. Rather than debating the theoretical impacts of policy ad nauseum or searching for clever means of empirically mimicking a randomized policy experiment, government should attack the problem directly. Just as random assignment of treatments is the most powerful method of testing for the causal

---

<sup>1</sup> *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](http://www.grazian-archive.com/archive/pdf/1965_06_00_ABS_salt_and_iron_debate_1_10.pdf).

<sup>2</sup> See Ian Ayres & Richard Brooks, *Does Affirmative Action Reduce the Number of Black Lawyers?*, 57 STAN. L. REV. 1807 (2005); Ian Ayres & John J. Donohue III, *Shooting Down the "More Guns, Less Crime" Hypothesis*, 55 STAN. L. REV. 1193 (2003); see also John J. Donohue III & Steven D. Levitt, *The Impact of Race on Policing and Arrests*, 44 J.L. & ECON. 367 (2001).

<sup>3</sup> Some terminology is helpful here. The experimental population is the group of all individuals taking part in an experiment. Some of the individuals in the population serve as experimental subjects—the individuals exposed to the experimental intervention. Other individuals in the experimental population serve as experimental controls. These individuals will be exposed to the existing policy or norm, and help the experimenter distinguish between the causal effects of the intervention and the background incidence of certain outcomes within a population that have nothing to do with the intervention.

impact of pharmaceuticals, randomly assigning individuals, firms, or jurisdictions to different legal rules can help resolve uncertainty about the consequences of laws and regulations.

The proposal is simple. Choose an area of policy debate. If the policy is a national one and applies to individuals, then randomly assign one policy to govern some individuals, while assigning the other policy to govern the actions of the other group. Next, observe what happens to the individuals assigned to different policies in the experiment. If the policy assignment has truly been random, then any differences in outcomes between the groups can be causally attributed to the differences in policy. Finally, apply the more effective policy to all individuals.

A similar process should be followed with respect to policies that must be applied to entire jurisdictions, such as gun control laws. If the policy debate concerns a policy that must be applied to an entire jurisdiction, then randomly assign different policies to different jurisdictions, observe the results, and apply the more effective policy to all jurisdictions.

The use of randomized experiments across states for the first time pays off on the idea that our federalist systems of independent states would create a “laboratory of democracy.”<sup>4</sup> To serve as laboratories for democracy, individual states could experiment with different laws and the rest of us could collectively sit back and learn from each other. But the problem has always been with the experimental design. As Susan Rose-Ackerman pointed out long ago, a good experiment requires a good control group.<sup>5</sup> The problem with many state-wide experiments is that they do not provide control groups. Alaska’s not really like Arizona. Similarly, the problem with allowing individuals to choose different policies is that individuals choosing one policy will be different from individuals choosing a different policy. In real labs, you don’t let the rats design the experiment. Instead, you randomize exposure to a treatment and observe the results. This is the exact content of our proposal. Widespread policy randomization offers the potential to provide answers to a host of policy debates.

Running policy experiments with subjects randomized to different treatments will not always be feasible. It would be hard to run a randomized experiment on whether to invade Iraq. Policy experiments may be extremely difficult to carry out from a pragmatic perspective as some may be extremely complex and some subjects may seek to change treatment groups. The British journal *BMJ* lampoons the insistence on randomized trials in evidence-based medicine rankings by pointing out that there has never been a randomized experiment testing whether parachutes actually work.<sup>6</sup> The parachute analysis underscores how similarly situated individuals may find it unfair to be randomly subjected to different regimes. If individuals or firms are competing, then differential treatment assignment may be more than unfair—it may undermine market dynamics.

In spite of these obstacles, some policy experiments have already proven incredibly useful. For example, Mexican President Ernesto Zedillo was the motivating

---

<sup>4</sup> *New State Ice Co. v. Liebmann*, 285 U.S. 262, 311 (1932). “It is one of the happy incidents of the federal system that a single courageous State may, if its citizens choose, serve as a laboratory; and try novel social and economic experiments without risk to the rest of the country.” *Id.*

<sup>5</sup> SUSAN ROSE-ACKERMAN, *CORRUPTION AND GOVERNMENT: CAUSES, CONSEQUENCES, AND REFORM* (1999).

<sup>6</sup> Gordon C.S. Smith & Jill P. Pell, *Parachute Use To Prevent Death and Major Trauma Related to Gravitational Challenge: Systematic Review of Randomised Controlled Trials*, 327 *BMJ* 1459-61 (2003).

force behind a randomized test analyzing the effectiveness of the Progresa Program for Education Health and Nutrition.<sup>7</sup> Under the Progresa program, the mothers of poor families were eligible for three years of cash grants and nutritional supplements if the children made regular visits to health clinics and attended school at least 85% of the time. To find out whether contingent subsidies worked, the experiment randomly assigned 506 villages to one of two groups: poor families in the Progresa villages were eligible for the subsidies (if they met the health and schooling contingencies); poor families in the non-Progresa villages just received the normal, pre-existing support. A group of international researchers then waited to see what happened to the 24,000 poor families living in both sets of villages. The results were dramatic. The teenagers in the Progresa villages went to school about a half year more during the initial two-year period. Anemia was down 12%, serious illness fell 12% and children in the Progresa villages were nearly a centimeter taller than their non-Progresa peers. From a public health perspective, an extra centimeter in less than three years is huge, because healthy kids grow faster.

The transparent power of these results has had a tremendous impact on policy both inside and outside of Mexico. In 2001, the program (now called “Oportunidades”) was expanded nationwide to cover more than two million families. Its budget in 2002 was \$2.6 billion or about 0.5% of the Mexican GDP. Today more than thirty countries have adopted Progresa-like contingent subsidy programs—including New York City, where Mayor Bloomberg has championed the privately funded Opportunity New York program.<sup>8</sup>

But randomized tests of public policy are also powerful for telling us what doesn’t work. The most important randomized policy experiment in recent years in the United States has been a \$70 million HUD-funded effort to find out what happens if poor families are given housing vouchers that can only be used in low-poverty (middle class) neighborhoods.<sup>9</sup> This “Move to Opportunity” (MTO) test randomly gave housing vouchers to very low-income families in five cities (Baltimore, Boston, Chicago, Los Angeles and New York City) and is collecting information for ten years on how vouchers impact everything from employment and school success to health and crime. The results aren’t all in yet, but the first returns suggest that there is no huge educational or crime-reduction benefit from moving poor kids to more affluent neighborhoods (with more affluent schools).<sup>10</sup> Girls who moved have been a little more successful in school and healthier, but boys who moved have done worse in school and are more likely to commit crimes. At the moment, it looks like the MTO data will overturn a widely held progressive belief that housing integration would mitigate many problems of the underclass.<sup>11</sup> Regardless of where the cards ultimately fall, the randomized MTO data is

---

<sup>7</sup> See Paul Gertler, *Do Conditional Cash Transfers Improve Child Health?: Evidence from PROGRESA’s Control Randomized Experiment*, 94 AM. ECON. REV. 336 (2004). The Progresa experience is also discussed at length in IAN AYRES, *SUPER CRUNCHERS: WHY THINKING-BY-NUMBERS IS THE NEW WAY TO BE SMART* (2007).

<sup>8</sup> See NYC CENTER FOR ECONOMIC OPPORTUNITY, *OPPORTUNITY NYC: CONDITIONAL CASH TRANSFERS* (2007), available at [http://home2.nyc.gov/html/ceo/downloads/pdf/report\\_opportunity\\_nyc.pdf](http://home2.nyc.gov/html/ceo/downloads/pdf/report_opportunity_nyc.pdf).

<sup>9</sup> John Goering, *Expanding Housing Choice and Integrating Neighborhoods: The MTO Experiment*, in *THE GEOGRAPHY OF OPPORTUNITY: RACE AND HOUSING CHOICE IN METROPOLITAN AMERICA* 127, 127-49 (XAVIER N. DE SOUZA BRIGGS & WILLIAM JULIUS WILSON EDS., 2005).

<sup>10</sup> Jeffrey R. Kling et al., *Experimental Analysis of Neighborhood Effects*, 75 *ECONOMETRICA* 83 (2007).

<sup>11</sup> OWEN M. FISS, *A WAY OUT: AMERICA’S GHETTOS AND THE LEGACY OF RACISM* (2003).

going to provide policymakers for the first time with very basic information about whether changing your neighborhood can change your life. And the contrasting results of the Progresa and MTO studies underscore that randomized methodology is politically neutral. It is not inclined to favor or disfavor a particular kind of intervention or non-intervention. The success of these experiments also underscores the enormous potential of widespread use of policy experiments for accessing the impact of statutes as well as regulations and other public policies.

Our proposal aims to systematize and expand randomized policy experiments to all areas of policy. Rather than implementing experiments such as Progresa or MTO on an ad hoc basis, government should view randomization and experimentation of policy as a core function. Just as the Office of Management and Budget's Circular Number A-94 mandates "benefit-cost analysis" of new regulations,<sup>12</sup> so too should new regulations and other policies require a randomization impact statement (RIS) that describes the causal impact of the policies estimated by randomized trials.

The remainder of this Article examines randomizing law and policy in greater detail. Part II presents a history of randomization and explains why randomization is such a powerful tool to test for causality. Part III lays out details of our proposal and addresses basic questions of how best to design randomized tests of law. Part IV discusses potential problems and pitfalls of randomized policy experiments, as well as responses to these complications. Part V concludes by suggesting how randomized experiments could inform three particular policy areas: securities law, civil rights law and tax law.

## II. The Power of Randomization

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, formally proposed using random assignments in agricultural trials in research growing out of his work at the Rothamsted Experimental Station.<sup>13</sup> In his 1935 book, *The Design of Experiments*,<sup>14</sup> 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."<sup>15</sup> 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."<sup>16</sup>

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:

---

<sup>12</sup> Office of Mgmt. & Budget, Circular A-94, Memorandum for Heads of Executive Departments and Establishments: Guidelines and Discount Rates for Benefit-Cost Analysis of Federal Programs (Oct. 29, 1992).

<sup>13</sup> RONALD A. FISHER, *STATISTICAL METHODS FOR RESEARCH WORKERS* (1925); *see also* R. Stevenson, Gold Standard for Drugs, *CHEMBYTES E-ZINE*, 9 (September). Parts of this section rely upon material also developed in AYRES, *SUPER CRUNCHERS*, *supra* note 7.

<sup>14</sup> RONALD A. FISHER, *THE DESIGN OF EXPERIMENTS* (1935).

<sup>15</sup> RONALD A. FISHER, *THE DESIGN OF EXPERIMENTS* 11 (6th ed. 1951).

<sup>16</sup> *Id.*

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.<sup>17</sup>

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 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, then 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 attributed. 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.<sup>18</sup>

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 didn’t take place until the late 1940s, when Austin Bradford Hill ran the first clinical trial testing the effectiveness of

---

<sup>17</sup> David P. Harrington, *The Randomized Clinical Trial*, 95 J. AM. STAT. ASSOC. 312 (2000).

<sup>18</sup> Fisher, *supra* note 15, at 18 (emphasis added).

streptomycin in treating tuberculosis.<sup>19</sup> By 1962, the use of random controlled trials had become so prevalent that Congress amended the Food, Drug and Cosmetic Act to mandate the use of “adequate and well-controlled investigations, including clinical investigations, by experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved.”<sup>20</sup> Since 1970, randomized clinical trials have been a critical part of FDA analysis of applications.<sup>21</sup>

And in the last twenty years the randomization tool has even spread to the testing of regulations themselves and government policymaking, as evidenced by the MTO and Progresá randomized experiments described in the introduction.<sup>22</sup>

### III. A Proposal of Presumptive Randomized Testing of Legal Change

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,<sup>23</sup> 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. Valid reasons for the absence of experimental data would include unambiguous need for a new policy, a pressing need for immediate action, the

---

<sup>19</sup> Medical Research Council, *Streptomycin in Tuberculosis Trials Committee: Streptomycin Treatment of Pulmonary Tuberculosis*, 2 BRIT. MED. J. 769 (1948).

<sup>20</sup> 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).

<sup>21</sup> 21 C.F.R. § 314.50. See *Abigail Alliance for Better Access to Developmental Drugs v. Von Eschenbach*, 445 F.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 Tort Reform*, 48 RUTGERS L. REV. 883 (1996); see also 40 C.F.R. § 799.9420 (EPA regulation mandating randomized testing of toxic substances).

<sup>22</sup> An Urban Institute study documented 143 randomized public policy experiments that had taken place by 1996. See DAVID GREENBER, DONNA LINKSZ & MARVIN MANDELL, *SOCIAL EXPERIMENTATION AND PUBLIC POLICY MAKING* 29 (2003).

<sup>23</sup> 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.

impossibility of conducting an experiment, and a de minimus exception. The norm, however, should be the presentation of data from a randomized policy experiment.

## **A. Choosing Policies To Test**

Conducting policy experiments imposes several costs and produces several benefits. The previous Part detailed the benefits, but remained silent on the costs. Because a policy's ripeness for experimentation depends on both the benefits and costs of experimental implementation, we must examine experimental costs in order to provide guidance for choosing policies for experimentation.

### **1. Costs of Policy Experimentation**

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. Policy experiments are "field experiments."<sup>24</sup> 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 but costly means of limiting such non-compliance. Even so, there will always be some number of non-compliers, and policymakers must ascertain means of preventing such non-compliance from biasing the results of the experiment.<sup>25</sup>

Finally, the costs and complexity of policy experimentation means that experimental "do-overs" will frequently be impossible; citizens may submit to an experiment once, but it is unlikely that they will do so again if the experimental design is flawed. For example, in 2006, the Woman's Health Initiative (WHI) reported the results of a \$415 million randomized test of the impact of a low-fat diet, which randomly assigned nearly 49,000 women ages 50 to 79 to follow, or not to follow, a "20% of

---

<sup>24</sup> For a discussion of field experiments in economics and public policy, see George Harrison & John List, *Field Experiments*, 42 J. ECON. LIT. 1009 (2004).

<sup>25</sup> For a detailed discussion of treating non-compliance to avoid bias, see Part IV.B.

calories from fat” diet and then followed their health for eight years.<sup>26</sup> The shocking news was that, contrary to prior accepted wisdom, the low-fat diet did not improve the women’s health. The women assigned to the low-fat diet weighed about the same and had the same rates of breast cancer, colon cancer, and heart disease as those whose diets were unchanged. But even supporters of randomized trials have argued that the study wasted hundreds of millions of dollars because it asked the wrong question. In retrospect, the critics argue that the study should have tested the impact not of a “20% calories from fat” diet, but a more severe 10% calories from fat diet. Dr. Michael Thun, who directs epidemiological research for the American Cancer Society, called the WHI study “the Rolls-Royce of studies” not just because it was high quality, but also because it was so expensive. “We usually have only one shot,” he said, “at a very large-scale trial on a particular issue.”<sup>27</sup> Examples such as the WHI study underscore that policy experiments must be done right the first time. To do this, extensive piloting and preparation is necessary, which adds to the already considerable costs of running such “field experiments.”

In addition to these implementation costs, experimental policies may have direct policy costs. That is, experimental policies may produce poorer outcomes than the incumbent policy. Indeed, that may be why the experimental policy has not been implemented in the first place. The costs born by the experimental subjects to a policy cannot be recouped if the experimental policy is disregarded.

## **2. Choosing Experimental Policy to Minimize Experimental Costs**

A systematic approach to randomized policy experimentation must address these costs. 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, other things 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

---

<sup>26</sup> Women’s Health Initiative, <http://www.nhlbi.nih.gov/whi> (last visited Sept. 3, 2008); *see also* Ayres, *supra* note 7; B.V. Howard et al., *Low-Fat Dietary Pattern and Weight Change Over 7 Years: The Women’s Health Initiative Dietary Modification Trial*, 295 JAMA 39 (2006); B.V. Howard et al., *Low-Fat Dietary Pattern and Risk of Cardiovascular Disease: The Women’s Health Initiative Randomized Controlled Dietary Modification Trial*, 295 JAMA 655 (2006).

<sup>27</sup> Gina Kolata, *Low-Fat Diet Does Not Cut Health Risks, Study Finds*, N.Y. TIMES, Feb. 8, 2006, at A1.

experiment with policies that have relatively positive expected effects.<sup>28</sup> 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.

Experiments should be limited to policies with potentially significant effects on behavior. Experimental implementation costs will often be somewhat fixed regardless of the potential impact of the policy. Conducting an experimental elimination of the penny is likely to cost nearly as much as conducting an experimental elimination of all denominations less than \$1. The former experiment's likely impact (positive or negative), however, is much lower. Even if the elimination of the penny proves to have surprisingly positive effects, it is unlikely that these effects justify the cost of the experiment. The answer to a larger scale question, by contrast, is more likely to yield information that justifies the experimental costs. Therefore, experiments should be limited to policies with relatively significant potential impacts.

Relatedly, 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 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*.

Experimentation will be cheaper when it is easier to convince the experimental subjects of the value of experimentation. We believe that repeals of existing restrictions will therefore be good candidates for experimentation. In cases such as the Sarbanes-Oxley (SOX) securities regulation restrictions, for example, many companies oppose the policy<sup>29</sup> but must comply with SOX's legal restrictions. Companies are likely to welcome a lottery that gives them the chance to avoid for a few years the compliance costs of SOX. An experimental repeal would give each company a chance to temporarily avoid SOX's restrictions, while current policy imposes SOX on everyone. Companies are therefore likely to support an experimental SOX repeal, making it an easy target for experimentation.

Finally, the costs of measuring outcomes will be lower when the desired outcome is straightforward. For example, if the purpose of eliminating the penny is to reduce transaction costs, then such costs can be measured relatively easily. If the purpose of the penny elimination is to increase aggregate national happiness, by contrast, then the costs of measuring the outcome of the experimental policy go up considerably. Other things equal, policies with straightforward and narrow outcome goals should be implemented before policies with broader aspirations.

---

<sup>28</sup> Yair Listokin, *Learning Through Policy Variation*, 118 YALE L.J. (forthcoming 2009) 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.

<sup>29</sup> See, e.g., U.S. CHAMBER OF COMMERCE, CAPITAL MARKETS, CORPORATE GOVERNANCE, AND THE FUTURE OF THE U.S. ECONOMY (2006), available at <http://www.uschamber.com/NR/ronlyres/eurumrodhxajii25owbcssgldn5qrcpr7abjwofcnj65gbc4a3ldm5vhbauvk7nkikezyyqj7yuxa73elm7udpiv7ga/060213capitalmarkets.pdf> (asserting that SOX's "excessive and unnecessary costs damage competitiveness and, ultimately, the interests of investors.").

### **3. Weighing the Costs and Benefits of Experimental Policies**

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. The causal effects of a new policy that is universally applied may be estimated by comparing outcomes before and after policy implementation. Such analysis is susceptible to the critique that the outcomes changed because of some other factor rather than the change in policy. Experiments, by contrast, alter the policy for some subjects but not others. If the subject and control groups are similar, then differences in outcomes cannot be attributed to other factors, since these factors should affect the subject and control groups to the same degree. Experiments therefore 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.

In total, experimental implementation of policy costs less than ordinary policy enactment and yields greater informational benefits. As a result, the threshold for experimental implementation of policy should be lower than the threshold for policy enactment.

### **4. Identification of Policies for Experimentation**

Even with a low threshold for experimentation, policymakers must still determine what policies to test from a vast universe of choices. At times, this will be easy. Some policies will be extremely controversial, with a significant but not unanimous portion of the community of people affected by a policy calling for change. Policies that are controversial because they have theoretically ambiguous and empirically uncertain effects are good candidates for tests involving randomization. But controversial policies are not the only candidates for testing, nor is every controversy the result of a genuine policy debate.

For example, some policies may greatly benefit a small group of individuals while slightly harming a much larger group.<sup>30</sup> The beneficiaries of such policies will not complain about them. Those harmed by the policy might, but the size of the harm may be so small that it is not worth the effort. In this case, a policy may be questionable in spite of the fact that no group criticizes the policy. Alternatively, a small group of potential beneficiaries may see benefits from changing a generally effective policy. This group may attempt to create a controversy where there is none by making noise or focusing on

---

<sup>30</sup> For a recent application of this oft-trodden ground of public choice theory where benefits are concentrated for one group, but costs are diffuse for another, see Paul Stancil, *Assessing Interest Groups: A Playing Field Approach*, 29 CARDOZO L. REV. 1273 (2008).

narrow failings of the policy rather than its overall effects. When this occurs, a policy may appear controversial in spite of its desirability.

It is important to emphasize, however, that these problems are less salient in the experimental context than in the ordinary policymaking context. If a group with concentrated benefits successfully lobbies for an experimental study of an ineffective policy, then the policy is likely to fail the experiment. Knowing this, lobbyists are unlikely to waste effort advocating for experimental policies that are likely to fail. Instead, they will focus their efforts on passing permanent policies, which lead to longer-lasting benefits for the few and are more difficult to empirically disparage. Lobbying for experimental studies will be limited to policies that have a fair chance of providing a successful outcome. And even if lobbyists occasionally push for experimentation of a failed policy, the costs of such experimentation are relatively low, as emphasized above.

As a result, we believe that policymaking bodies should have discretion when choosing policies for experimental randomization. Academic research, practitioner opinion, and citizen opinion should all be inputs into the decision regarding policies for testing. These inputs cannot be decisive, however. There is no substitute for the judgment of the policymaker in deciding what policies to test.

In the legislative context, this discretion should be unconstrained. Any policy that receives majority support for an experimental study likely passes the plausibility test required for reasonable experimentation. Requiring additional justification for experimental implementation is not worth the cost.

Policies implemented as experiments in administrative agencies, by contrast, will generally require a smaller number of supporters than legislative experiments. As a result, some limitation of experimental policy discretion is justified in order to limit experiments to policies worth trying. Given the high value and low cost of experiments, however, the limitations should be minimally intrusive. To constrain the discretion of the policymaker and make certain that the inception of a randomized test is legitimate, the administrative agency should include a description of how it came upon a policy for testing in the randomization impact statement.

Even if policymakers have agreed on experimenting in a policy area, they must still choose between potentially infinite variations on a given policy. If policymakers are considering an experimental repeal of the Sarbanes-Oxley securities law, for example, they must consider which of SOX's many provisions to repeal. Some of these provisions may be controversial, but others may enjoy widespread support. If the entire law is experimentally repealed, then the estimated impacts of the law will confound the repeal of the contested provisions with the repeal of the clearly desirable ones. This will both raise the cost of the experiment—by eliminating desirable provisions—and reduce the informational value of the experiment—by confounding different effects.

Just as discretion should be the rule in choosing experiments at the broad policy level, so too should policymakers have discretion when choosing between variations of policies that share the same basic structure. Several rules of thumb apply to guide these decisions, however. First, policymakers should of course avoid variations that are widely agreed to be less effective than alternative variations. Second, policymakers should attempt to implement a “mainstream” variation, to prevent arguments that an experimental policy failed because of the unusual variation rather than a flaw in the broader policy mechanism.

Finally, if the added complexity does not prove too daunting, policymakers can simultaneously test multiple dimensions of legal change. The experimentation literature has developed sophisticated methods (including fractional factorial and Taguchi methods) to parsimoniously explore the impact of changing different dimensions of an advertisement or business promotion.<sup>31</sup> It is possible, for example, to simultaneously test changes to SOX that vary: (i) the frequency of required reporting; (ii) the content of reporting requirements; and (iii) the penalties for failure to report. A well-designed experiment can study a fraction of the possible policy permutation and make predictions about how untested permutation would fare.<sup>32</sup>

## ***B. Experimenting with Policies***

Once a particular policy has been chosen for experimentation, the experiment must be implemented. Policymakers must choose both the level and scope of the randomization and then randomly assign experimental populations to the treatment or control groups. This Section examines how such decisions should be made.

### **1. Choosing the Unit of Randomization and Experimentation**

A policy can be randomly assigned at many different levels of randomization. Some policies can be randomly assigned at the individual level. This level of randomization is familiar from the pharmaceutical industry. In a drug trial, some individual subjects are given the experimental drug, while other individuals 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, the infamous "Part D," randomly assigned more than six million people to one of up to twenty qualified state plans.<sup>33</sup> 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.

---

<sup>31</sup> See AYRES, *supra* note 7 (describing Taguchi experimental methods) and JOHANNES LEDOLTER & ART SWERSY, TESTING 1-2-3: EXPERIMENTAL DESIGN WITH APPLICATIONS IN MARKETING AND SERVICE OPERATIONS (2007).

<sup>32</sup> LEDOLTER & SWERSY, *supra* note 31.

<sup>33</sup> RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS (2008).

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.<sup>34</sup> Thus, random assignment to individuals would be the best strategy when a policy targets individual outcomes and there are no spillovers to (untreated) other individuals.<sup>35</sup> 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,<sup>36</sup> 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.

---

<sup>34</sup> When policy is randomized at the state level, for example, serial correlation in error terms makes standard errors wide, and therefore complicates the finding of statistically significant policy impacts. For details, see Marianne Bertrand, Esther Duflo & Sendhil Mullainathan, *How Much Should We Trust Differences-in-Differences Estimates?*, 119 Q. J. ECON. 249 (2004).

<sup>35</sup> Relate to the unit of observation literature in economics?

<sup>36</sup> 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).

## 2. Choosing the Experimental Subjects

After choosing the level of randomization for an experimental policy, policymakers must choose the appropriate scope for the experiment. That is, policymakers must decide who should be included in the pool of individuals that randomly receive the treatment policy or a control policy. As with choosing the level of randomization, there are many possibilities. Policymakers could induce individuals, firms, or jurisdictions to volunteer for programs using financial incentives or the possibility of avoiding unpopular restrictions. Alternatively, policymakers could compel experimental units to take part in experiments. Even if policymakers choose the path of compulsion, they must decide how large to make an experiment. They could compel all possible units to take part in an experiment, or only compel a subset of these units to participate.

We believe that volunteer experimental participation is best when volunteers provide a representative estimate of the causal effects of a policy. If volunteer randomization provides a skewed estimate of the relevant effects, then policymakers should use compulsion to obtain a representative experimental size. When using compulsion, however, policymakers should pick the smallest possible experimental size that will yield accurate estimates of the causal effects of the experimental policy.

From a moral perspective, voluntary participation in experiments is preferable to compelled participation.<sup>37</sup> Instead of invoking the power of the state to conduct experiments, relying on voluntary participation allows private self-interest to create the experimental pool. Indeed, the most familiar body of randomized experimentation, drug testing, relies on volunteers (sometimes with financial incentives and sometimes without). Therefore, involving only volunteers in policy experiments should be the initial aim of all experiments, whatever the unit of randomization. If jurisdictions or firms are the experimental unit, then the decision-making bodies of such units can “volunteer” their organizations for an experiment.

Unfortunately, volunteers sometimes provide unhelpful estimates of the causal effects of policy.<sup>38</sup> 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.<sup>39</sup> 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. If all volunteers for an experimental chemotherapy drug have cancer, then the experiment will give a very poor estimate of the average causal effect of giving the drug to all individuals. If the volunteers respond to the experimental drug in a representative way for cancer patients, however, then the volunteers provide a good estimate of the causal effect of the drug on cancer patients.

---

<sup>37</sup> Voluntary participation includes both traditional volunteering, as well as offering to participate in an experiment in return for financial incentives.

<sup>38</sup> A related problem of interpretation—how to obtain causal effects estimates when some individuals drop out of an experiment but others do not—is discussed in Section IV.B, *infra*.

<sup>39</sup> 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).

The drug volunteers, however, may not even provide good estimates for the effect of the experimental chemotherapy on cancer patients. Volunteers are not like everyone else. Some of them may be sicker than the average cancer patient and therefore ready to try unproven therapies. The effect of the experimental drug on particularly sick cancer patients may be different than the effect of the drug on the average cancer patient. While experimenters can attempt to control for the sickness of each patient, their controls are likely to be imperfect. In total, the volunteers must be representative of the group of people targeted by the drug for the experiment to provide valuable estimates of causal effects of the treatment on the general population. Or stated slightly differently, the volunteer experiment shows the causal effect of the treatment on the type of person who is willing to volunteer, but may not provide evidence of how the treatment would affect those who would not be willing to volunteer.

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. Although experimenters can attempt to control for differential effects (e.g., people with high skill levels are less likely to be helped by the program than those with fewer skills, but experimenters can collect data on existing skills and control for this effect), some of the variables that affect the response to the job skills program for volunteers will be unobservable. For example, 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 therefore biased. The experimental effect estimate of the program using volunteers, however, is an unbiased estimate of the effect of making the program voluntary but expanding its availability. Voluntary experiments thus first and foremost can guide policymakers on whether a particular policy should be offered to the general population. Under voluntary programs, the government's offer is in some sense the treatment. Further inferences about mandatory policies again are only as valid to the extent that the volunteers are representative of those who did not volunteer.<sup>40</sup>

This analysis has several implications for experimental design. First, policymakers must think long and hard about the targets of the policy when designing their experiments. Failure to do so may result in inappropriate experimental effect estimates. Second, when the volunteers for an experimental policy do not represent those who will be affected by the new policy (e.g., the government is considering mandating the job skills program for a group of people), then policymakers may have to compel participation in the experiment for a group of people that is representative of the group of people that will be targeted by the eventual policy. Because such compulsion is costly,

---

<sup>40</sup> Ultimately, policymakers are interested in "policy-relevant treatment effects," which are the effects of a policy on the people who will be affected by the policy. See James J. Heckman & Edward Vytlacil, *Policy-Relevant Treatment Effects*, 91 AM. ECON. REV. 107, 107-08 (2001).

the size of the group compelled to participate in an experiment should be the smallest size consistent with obtaining robust estimates of the causal effects of the policy.

Compelling individuals to take part in policy experiments appears morally problematic, an issue we address in Part IV.D.

### 3. Experimental Length

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 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 result, it may be appropriate to shut down the study earlier than expected.<sup>41</sup> 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 the 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.<sup>42</sup>

Once the level of randomization and the experimental population (compelled or volunteer) of the randomized policy experiment are chosen, the experiment must be carried out. Rather than having each policymaking body administer its own experiments, we believe it will be more efficient to have a single agency that specializes in administering experiments designed by policymakers. Running policy experiments requires specific skills, such as knowing what types of outcome information are readily

---

<sup>41</sup> 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.

<sup>42</sup>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).

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. If every policymaking body conducts many policy experiments, then each body may develop its own expertise. In the early stages, however, the policymaking bodies are likely to have a steep learning curve for running experiments. As a result, a specialized experimental administration agency will generally be preferable to having each policymaking body administer its own experiments. Just as pharmaceutical companies hire clinical trial companies to run drug trials, so too should policymaking bodies use experimental trial specialists.<sup>43</sup> Even if a governmental experimental agency is not feasible, private sector organizations specializing in experiments provide an alternative to each policymaker going it alone on all experiments.

The raw data from policy experiments should be made available to the public. This will allow academic researchers and other interested parties to evaluate the policies in addition to the policymaking body sponsoring the experiment. Public availability allows conclusions about the effect of experimental policies to be appropriately tested and debated.

### ***C. Enacting Permanent Policies***

While randomized experiments provide information about the likely outcomes of a policy, they cannot yield conclusions about the desirability of various outcomes. Policymaking bodies must therefore use discretion when determining what experimental policies should be enacted as permanent laws or rules. The discretion, however, should be informed by the results of the policy experiments. If the policy is enacted as a statute, then public debate on the statute should reference the results of the experiment. If the policy is enacted as a rule, the rulemaking body should issue a randomization impact statement (RIS) for any proposed new rule.

An RIS should include the following elements (some of which are discussed above):

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)<sup>44</sup> to resolve the legislative debate, because legislators and their constituent groups may have incommensurable preferences.<sup>45</sup>

---

<sup>43</sup> 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).

<sup>44</sup> *See Planned Parenthood of S.E. Pa. v. Casey*, 505 U.S. 833 (1992) (striking down Pennsylvania spousal notification law). But a state might experiment on offering the possibility of giving couples at the time of

2. A detailed description of the experiment. The description should discuss the unit of randomization, the scope and length of the experiment, and the possible effect 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.

The RIS should be submitted by the agency to the Office of Management and Budget (OMB), just as agencies currently submit cost benefit analyses to the OMB under Executive Order 12,866.<sup>46</sup> The RIS will also serve as important source documents, allowing Congress and even the courts to oversee agencies.

Although not required, RIS analogues from sponsors of legislation would be extremely helpful in framing debate and discussion. For example, policy sponsors seeking to forego experimentation should explain why their proposed policies should be enacted as permanent policies rather than experimental policies. If there has already been an experiment, then an RIS will provide a useful blueprint for the arguments in favor and against the policy in light of the information provided by the experiment. In total, the RIS framework used in conjunction with systematic policy randomization offers the prospect of a significant improvement in the policymaking environment.

## IV. Caveats

While we believe the previous Part presented a formidable case for systematic policy randomization, experimental randomizations present their own set of difficulties. Indeed, some of these difficulties, such as the complexity of finding relevant policy effect estimates from randomized experiments, have already been examined. This Part examines some further impediments to randomized policy experimentation and suggests possible responses.

---

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).

<sup>45</sup> 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 unaborted 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.

<sup>46</sup> Exec. Order No. 12,866, 58 *Fed. Reg.* 51,735 (Sept. 30, 1993). Indeed, the experimental data discussed in the RIS is likely to inform the cost benefit analysis required under this order. As a result, the RIS requirement is symbiotic with the cost benefit analysis requirement.

## ***A. Randomized Experiments that Are Not Double-Blind***

The purest form of randomized experiments also includes informational control on both the researcher and the subjects. In double-blind experiments, for example, neither the researcher 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 can't favor a particular outcome. Hence, double-blind designs can protect against "observer bias."<sup>47</sup>

Keeping subjects in the dark as to whether they are in the treatment group or not analogously insures that their behavior and emotional outlook is 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.<sup>48</sup>

Even when subjects don't know whether they are in the treatment or control group, they will often know that they are participating in a randomized experiment and the participation knowledge itself may impact the outcomes. The impact of knowing that they are being observed might, for example, make subjects alter their behavior to (dis)please the researcher. In 1955 Henry Landsberger coined the term "Hawthorne effect" to refer to effects induced by subject knowledge of the experiment.<sup>49</sup> The name came from a set of ergonomic experiments ran at the Hawthorne Works near Chicago. The researchers found a short-term improvement in worker performance after almost any change in lighting.<sup>50</sup> But productivity soon returned to normal levels. While somewhat ambiguously defined (because workers had knowledge of both the experiment and when they were in treatment group), the idea of Hawthorne effect has come to refer to the impact of observing on the observed. 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.<sup>51</sup>

In randomized tests on laws and public information, it will be ever harder to keep subjects in the dark about how they are being treated or that they are subjects in an experiment. For example, we will discuss below an SEC experiment that might repeal certain Sarbanes-Oxley reporting requirements for randomly selected corporations. It would be impossible not to let the corporations know whether or not they were subject to the reporting requirement. But the transparency of the reporting requirement is not as large a concern here. In the medical arena, researchers are primarily interested in the

---

<sup>47</sup> RON MCQUEEN & CHRISTINA KNUSSEN, INTRODUCTION TO RESEARCH METHODS AND STATISTICS IN PSYCHOLOGY (2006).

<sup>48</sup> 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).

<sup>49</sup> HENRY A. LANDSBERGER, HAWTHORNE REVISITED (1958).

<sup>50</sup> RICHARD GILLESPIE, MANUFACTURING KNOWLEDGE: A HISTORY OF THE HAWTHORNE EXPERIMENTS (1985).

<sup>51</sup> 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).

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 you are treated becomes part of the treatment, but this is not a bad thing, because the researcher wants to know whether a known legal change will have an impact.

However, as before, there is still a separate Hawthorne concern. While knowledge of the legal change is desirable, subject knowledge that they are being observed as part of a test may not be desirable if the goal is to produce information about whether to make a legal change (where the population will no longer be subjects in an ongoing policy experiment). But knowledge of the legal change does not necessitate that subjects know that they are taking part in a randomized study. For example, one could imagine a test of speed limits 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. Concerns over Hawthorne effects will militate toward longer series of ongoing tests to determine whether impacts of laws are long-lived or merely the products of observation upon the observed.<sup>52</sup>

---

<sup>52</sup> Anup Malani has developed a clever mechanism for estimating the size of placebo effects when subjects know that they are being tested but not whether they are part of the treatment or control group. Anup Malani, *Identifying Placebo Effects with Data from Clinical Trials*, 114 J. POL. ECON. 236 (2006). By manipulating the known probability of being in the control group, Malani was able to compare the pure information effect. For example, he estimated the placebo effect of statins by running two different clinical trials—one in which there was a 50% chance of being treated with statins (and a 50% chance of receiving a placebo) and one in which there was a 100% chance of being treated with statins. He then looked at the average reduction in LDL cholesterol only among the people in each trial that actually received the statin. By holding the drug treatment constant and only varying the subjects' probabilistic knowledge of receiving the treatment, he could measure the independent effect of the information, which in this case increased the size of the LDL reduction by a statistically significant 28% (the probability of "nocebos" side effects also increased by almost 50%). Milani's method might also be applied to tests of regulatory change. For example, if the legal change concerns the effort that the IRS will make to enforce a law, subjects might just be told that there are different probabilities of audit and look for differences among those who are actually audited.

## ***B. Difficulties of Estimating Policy Relevant Treatment Effects from Randomized Experiments***

The last subsection concerned informational problems in control that can potentially undermine the ability to extrapolate the results from a randomized policy experiment to the general population. But analogous problems of extrapolation also apply to other dimensions of testing. In the prior section, the extrapolation was in going from a sample with certain informational attributes—such as subjects knowing they were participating in an experiment—to a population with different informational attributes. In this section, we will consider other attributes besides information in which 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.”<sup>53</sup> For example, as discussed above,<sup>54</sup> 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. So if a job search assistance program for the unemployed is shown in a randomized trial of volunteers to provide an increase in employment, the natural policy response would be to *offer* the job search assistant program more generally. It would not be to mandate job search

---

<sup>53</sup> 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); James Heckman, Jeffrey Smith & Christopher Taber, *Accounting for Dropouts in Evaluations of Social Experiment* (Nat'l Bureau of Econ. Research, Working Paper No. 166, 1994); James Heckman, *Alternative Approaches to the Evaluation of Social Programs: Econometric and Experimental Methods*, Address at the World Congress of the Econometric Society (1990); James Heckman & Jeffrey Smith, *Evaluating the Welfare State*, Address at the Frisch Symposium (March 1995); James Heckman & Jeffrey Smith, *Ashenfelter's Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies* (1994) (unpublished paper, University of Chicago); James Heckman, *The Case for Simple Estimators: Experimental Evidence from the National JTPA Study* (1993) (unpublished paper, University of Chicago); James Heckman & Rebecca Roselius, *Evaluating the Impact of Training on the Earnings and Labor Force Status of Young Women: Better Data Help A Lot*, (1994) (unpublished paper, University of Chicago); James Heckman, Hidehiko Ichimura, Jeffrey Smith & Petra Todd, *Non-Parametric Characterization of Selection Bias Using Experimental Data: A Study of Adult Males in JTPA* (1995) (unpublished paper, University of Chicago); James Heckman, *Randomization as a Multiple Instrumental Variable*, (1993) (unpublished paper, University of Chicago); James Heckman & Jeffrey Smith, *Substitution Bias in Social Experiments: An Analysis of the JTPA Data* (1994) (unpublished paper, University of Chicago).

<sup>54</sup> See Section I.B.2, *supra*.

assistance for all the unemployed. One would need a separate theory or analysis of representativeness of the volunteers to make an inference about a mandated program.

Still, it is useful to note by way of comparison how routinely in medicine one moves from randomized tests on volunteers to quasi-mandatory across-the-board treatment proposals. Moreover, as noted above, government can respond to this “voluntariness” problem by designing tests with mandatory participation. Ethical rules require that patients consent to participation in medical RTCs, 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.<sup>55</sup>

Another extrapolation difficulty concerns attrition of subjects during the course of the randomized experiments. In many RTCs, it is inevitable that some proportion of the subjects will disappear and/or refuse to cooperate with further data collection during the course of the RTC itself. Researchers by definition are unable to observe the outcomes of interest with regard to these missing observations. For example, if the government wants to test the impact of housing vouchers on crime, some number of the subjects in both the control and treatment groups will drop out of the sample. Attrition of this kind can bias the results because the average outcomes of those subjects who remain in the study to the end can differ from those who drop out along the way.<sup>56</sup> Success among those that are observed is not necessarily indicative of success among those that are not observed. For example, Dean Karlan and Jonathan Zinman analyzed the impact of smoking cessation commitment bonds in the Philippines where some subjects posted monetary bonds (averaging about \$11) that would be returned if they passed a nicotine test six months later. An obvious concern is that subjects who continued to smoke would drop out, not bothering to take a test that they were sure to fail. Some of the authors’ analysis responded to this concern by coding all attrition as failures. More generally, there is abundant evidence that well-designed randomized studies can produce low rates of attrition.<sup>57</sup> The government in particular can put in place carrot-and-stick incentives to keep subjects participating. Subjects who want to remain eligible for government services will be loathe not to respond when the Leviathan calls.

### **C. The Bias of Incomplete Knowledge**

When the tested subjects are representative of the larger population, there will not be an extrapolation concern. But even when a valid causal inference can be made about

---

<sup>55</sup> 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).

<sup>56</sup> JERRY A. HAUSMAN & DAVID A. WISE, SOCIAL EXPERIMENTATION (1985).

<sup>57</sup> E.g., Steven E. Hearne, et al., *Internal Mammary Artery Graft Angioplasty: Acute and Long-Term Outcome*, 44 CATHETERIZATION AND CARDIOVASCULAR DIAGNOSIS 153 (1998); Larry C. Lyons & Paul J. Woods, *The Efficacy of Rational-Emotive Therapy: A Quantitative Review of the Outcome Research*, 11 CLINICAL PSYCHOLOGY REV. 357 (1991).

the impacts of the treated law relative to a control, policymakers will invariably have questions about how untested policies would have fared. As discussed above, the cost and time to test will limit the number of policies tested. Particularly when results are limited to dichotomous “champion/challenger” tests, analysts will have no information about alternatives. Ideological advocates of particular policies will not go quietly into the night. After the fact, they will argue that the wrong “challenger” was tested.

This untested policy problem is particularly a concern with regard to questions of how to optimally tailor laws and regulations. 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.<sup>58</sup> For example, imagine that a speed limit study randomizing across different cities shows that twenty mph limits produce *more* accidents than thirty mph limits—or more particularly that the average number of accidents in the twenty mph group was higher than the average accidents observed in the thirty mph group. Policy makers would still not know how twenty-five or thirty-five mph limits would fare. But they would also not have known—without more analysis—whether twenty mph limits were not in fact optimal for certain types of cities. It might still be, even though twenty mph limits on average produced more accidents, that small, rural cities fare better with the lower limit. It is possible to run regressions on the results of randomized studies to test to see if the average result holds true for subgroups within the tested sample. But causal inferences about the impact of policies on subgroups are less straightforward, and like other regression analyses, turn on the inclusion and exclusion of adequate control variables.

It is possible to use randomized testing to independently test whether tailored and/or contingent policies outperform untailored policies. Thus, subsequent speeding trials could be conducted to test whether a tailored policy (of twenty mph in small cities and thirty mph in large cities) is superior to an untailored policy (of thirty mph in all cities).<sup>59</sup> But the possibilities for tailoring in any particular arena are endless and it is unreasonable to expect that more than a tiny fraction 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.

#### ***D. The (Im)Morality and (Un)Constitutionality of Randomized Experiments***

Finally, even if randomized tests can produce valid causal inferences about which law is best among a rich set of potential contenders, there is a final issue about whether it

---

<sup>58</sup> James J. Heckman, *Detecting Discrimination*, 12 J. ECON. PERSPECTIVES 101 (1998).

<sup>59</sup> 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](http://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).

is moral to conduct randomized experiments of law. To randomly subject citizens to different laws smacks of capriciousness and even raises questions of whether randomized tests of legal rules would be constitutional under the Due Process Clause. Indeed, Justice Potter Stewart in *Furman v. Georgia* criticized a state's criminal system because:

For, 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.<sup>60</sup>

But in thinking about the Due Process Clause's prohibition on capriciousness, it is important to distinguish random and non-random, but arbitrary, operation of the law. Indeed, Justice Marshall in *Furman* concluded that "It also is evident that the burden of capital punishment falls upon the poor, the ignorant, and the underprivileged members of society."<sup>61</sup> But in some sense, Justices Stewart and Marshall can't both be right. If Marshall was correct (and there is abundant evidence that he was),<sup>62</sup> that (controlling for reprehensibility) 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,<sup>63</sup> 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 (ex post the realization of the coin flip) arbitrarily treating equal people different.

Justice O'Connor, in *Ohio Adult Parole Authority v. Woodard*,<sup>64</sup> 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."<sup>65</sup>

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

---

<sup>60</sup> *Furman v. Georgia*, 408 U.S. 238, 313 (1972). *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.").

<sup>61</sup> *Id.* at 365-66 (Marshall, J., concurring).

<sup>62</sup> DAVID C. BALDUS, ET AL., *EQUAL JUSTICE AND THE DEATH PENALTY: A LEGAL AND EMPIRICAL ANALYSIS* (1990).

<sup>63</sup> *Stevens v. Marks*, 383 U.S. 234 (1966).

<sup>64</sup> 523 U.S. 272, 288 (1998).

<sup>65</sup> *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 . . .").

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.<sup>66</sup> 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.<sup>67</sup> 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.<sup>68</sup> 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.<sup>69</sup>

While many people are viscerally appalled by the notion of judges flipping coins to decide legal issues, coin flipping need not be a "meaningless ritual." In particular contexts, there are a variety of public policy rationales for randomized decisions. It'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.<sup>70</sup> 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.<sup>71</sup> (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.<sup>72</sup> 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

---

<sup>66</sup> *People v. Jones* (N.Y. Crim. Ct. 1982).

<sup>67</sup> Gary Slapper, *Weird Cases: Justice by Coin-Toss*, TIMES ONLINE, Nov. 16, 2007, <http://business.timesonline.co.uk/tol/business/law/article2882090.ece>.

<sup>68</sup> *Id.*

<sup>69</sup> 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](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>.

<sup>70</sup> Ian Ayres & Eric Talley, *Solomonic Bargaining: Dividing a Legal Entitlement to Facilitate Coasean Trade*, 104 YALE L.J. 1027 (1995).

<sup>71</sup> 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).

<sup>72</sup> JON ELSTER, *SOLOMONIC JUDGMENTS: STUDIES IN THE LIMITATION OF RATIONALITY* (1989).

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.<sup>73</sup> 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. (This deterrence result is, however, reversed for risk-preferring criminals, and it is troubling that Judge Jones before flipping did not inquire of the defendant if he was a betting man.)<sup>74</sup>

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.<sup>75</sup> By now, you should be able to see how a randomized study structured at the individual or jurisdictional level could answer this dispute. Justice O’Connor is appalled by the idea of clemency by chance. But even here we think there would be value in randomly granting clemency and parole to those inmates who just didn’t make the cut, to see if in fact they had a higher recidivism rate than those who did. (And randomly reducing the harshness of sentences appears less “cruel and unusual” than random enhancements from some status quo benchmark.)

Stepping back, this informational rationale for randomization is at the heart of this Article. The idea that government programs and protections would at random be denied to some citizens or that government mandates would randomly be imposed on some citizens seems capricious. But the great justification for randomized application of law is to educate lawmakers and citizens alike about which policies work. The informational rationale also acts as a principle for deciding when not to test and when to stop testing. We shouldn’t allow randomized tests of parachutes,<sup>76</sup> 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.<sup>77</sup>

Still, the idea of legal experimentation on citizen subjects raises the Kantian difficulty that the process seems to treat citizens merely “as a means to an end” and not as independent ends in and of themselves.<sup>78</sup> Medical protocols may in part resolve this difficulty by requiring that participating subjects be given adequate information to give informed consent.<sup>79</sup> But as discussed above, informed consent can weaken the power of tests if the sample of consenting participants is unrepresentative of the larger population

---

<sup>73</sup> Steven Shavell, *Economic Analysis of Public Law Enforcement and Criminal Law* (Nat’l Bureau of Econ. Research, Working Paper No. 9698, 2003).

<sup>74</sup> Editorial, *For Whom the Coin is Tossed*, N.Y. TIMES, February 13, 1982, at 24.

<sup>75</sup> Jeffrey R. Kling et al., *Experimental Analysis of Neighborhood Effects*, 73 ECONOMETRICA 83 (2007).

<sup>76</sup> See *supra* note 7.

<sup>77</sup> 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).

<sup>78</sup> “Act in such a way that you treat humanity, whether in your own person or in the person of any other, always at the same time as an end and never merely as a means to an end.” IMMANUEL KANT, METAPHYSICS OF MORALS (DATE).

<sup>79</sup> WORLD MEDICAL ASSOCIATION GENERAL ASSEMBLY, WORLD MEDICAL ASSOCIATION DECLARATION OF HELSINKI: ETHICAL PRINCIPALS FOR MEDICAL RESEARCH INVOLVING HUMAN SUBJECTS ¶ 29 (2002), available at <http://www.wma.net/e/policy/pdf/17c.pdf>; S.J.L. Edwards, R.J. Lilford, J.C. Jackson, J. Hewison, J. Thornton, *Ethical issues in the Design and Conduct of Randomised Controlled Trials*, HEALTH TECH. ASSESSMENT (forthcoming).

of interest. The more information that is provided—especially new participants over the course of ongoing trials—the more severe this participation problem is likely to be.

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.<sup>80</sup> 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.<sup>81</sup> Notwithstanding the supposed requirements of informed consent, medical trials routinely fail to give participants the best current information about the probable winner in a champion/challenger trial.<sup>82</sup> The reason for the failure is keep patients participating. Would you want to participate in a randomized study where one of the therapies had a 70% chance of being more effective? Patient surveys indicate, unsurprisingly, that “[w]illingness to undergo randomisation drops as prospective participants are given more preliminary data and as they are made aware of any accumulating evidence of effectiveness.”<sup>83</sup>

In many ways, however, the difficulties of running randomized trials on the law raises less serious concerns than medical trials, where the consequences may entail the irreversible loss of life. In a representative democracy, the informed consent of elected representatives might at least in part substitute for the informed consent of individuals. After all, legislatures without individualized consent have since time past memory been running non-randomized legal experiments—for example, passing temporary statutes with automatic sunset provisions. It is better to establish a regulatory structure and legislative norms that limit randomized experiments to circumstances where there is genuine dissensus about whether passing or repealing a law would enhance social welfare.

### ***E. The Political Economy of Randomized Testing***

Even if randomization offers considerable benefits, policymakers may not have the appropriate incentives to implement systematic experimentation of controversial policies. Few businesses, for example, have undertaken systematic experimentation.<sup>84</sup> If

---

<sup>80</sup> 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).

<sup>81</sup> R.J. Lilford & J. Jackson, *Equipoise and the Ethics of Randomisation*, 88 *J. ROYAL SOC’Y MED.* 552 (1995).

<sup>82</sup> 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.”).

<sup>83</sup> 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.

<sup>84</sup> One of us has consulted with several Fortune 500 companies that have never, to date, run a randomized test. And while many sports teams have caught the Moneyball train and started data mining of historical

businesses have yet to adopt widespread randomization, can there be any hope for widespread adoption by government? We believe the answer is yes, as the MTO, Progres, and other examples suggest. Indeed, it is at least arguable that government has been less resistant to the randomization technique than business.

How could this be? Usually we might expect business to be more adroit than government in picking up on profitable technologies. Why would business have a structural advantage in adopting this testing methodology? We tentatively suggest that the two-party division of government may perversely have made randomization more politically feasible than the more unitary, top-down structure of American business. In business, randomization is only likely to occur with active support from the CEO. But many CEOs still show unwillingness to put their intuitions to the test. Companies are more likely to run corporate “experiments” where the management hand-picks the sample to try out a new idea. But the problem is that these non-random experiments have no obvious control group to compare.

In contrast, politicians’ ideological differences often lead them to believe their side will “win” the experiment. If you think that government interventions are generally not worth the cost, you’ll likely think that an experimental intervention of Headstart will fail. If you’re presumptively pro-intervention, you’ll tend to think it’ll succeed.<sup>85</sup> Hence, ideological optimism inflates both sides’ senses that the results will ultimately vindicate their policy preferences.

Politicians are also willing to take a chance because they trust the results much more than the results of regression. Do concealed weapons laws increase or reduce crime? Does the death penalty deter crime? Politicians have pointed to competing regression studies to support their preferred predilections. It is relatively easy to dismiss regressions that you don’t like by merely arguing that the studies don’t sufficiently control for extraneous variables that you believe are driving the results.<sup>86</sup> In contrast, it is much harder to dismiss the results of randomized experiments. On the Internet, for example, if 100,000 people see an insurance ad with a caveman and another 100,000 at random are chosen to see an ad with a gecko, simply comparing the average buy rate among the two groups provides powerful evidence of which ad is more effective. The same holds true for randomized tests of public policy. Politicians are more likely to abide by the results of a randomized study than by the results of an election. At the end of the day, politicians have to trust the statistician to have done the regression correctly.

Randomizing testing is also a perfect alternative for those few elective officials who don’t have an ideological axe to grind and simply want to know whether a bright

---

results, there has yet to be a random test of sports strategy. (In a 2007 *New York Times Freakonomics* post, one of us challenged readers to come up with a single example of a randomized test of sports strategy. To date, the best example comes from the legal polymath, Pam Karlan, who pointed out that in eight-person shells some crew coaches use “seat races” where they switch individual pairs to assess the strength of individual athletes.) It would be straightforward to run an RTC to see: i) whether a distraction technique reduces an opponent’s free throw percentage; and ii) whether a statistical algorithm did a better job of calling pitches than a manager or catcher. It would be difficult to imagine medical research not exploiting both regression and randomization, but thus far the dramatic movement toward evidence-based sports strategy to date has only exploited the mining of historic data.

<sup>85</sup> John J. Donohue III, *Guns, Crime, and the Impact of State Right-to-Carry Laws*, 73 *FORDHAM L. REV.* 623 (2004).

<sup>86</sup> Dan M. Kahan, *Social Meaning and the Economic Analysis of Crime*, 27 *J. LEGAL STUD.* 609 (1998).

idea actually works. Randomized testing in government offers conflicted politicians the perfect compromise, where they can put off making the hard substantive decision initially (and even later can defer to the results of the study).

## V. Applications

The previous Parts laid the theoretical groundwork for systematic randomized policy experiments and briefly described some policy experiments. In this Part, we examine randomized policy experiments in action. To demonstrate the generality of such experiments, we develop a number of policy experiment applications to different fields of law and policy. We begin with a detailed examination of a real-life randomized securities law experiment.

### A. 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.<sup>87</sup> 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.<sup>88</sup> In total, 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.

#### 1. 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.

The SEC described the short sale restrictions as follows:

Short selling in exchange-listed stocks (“Listed Stocks”) in the U.S. has been subject to a “tick test” since 1938. Rule 10a-1 under the Securities Exchange Act of 1934 allows short sales to occur only at an uptick or a zero uptick (also known as a “zero-plus tick”) for Listed Stocks. That is, short sales in Listed Stocks may

---

<sup>87</sup> 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).

<sup>88</sup> 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).

be effected above the last trade price or at the last trade price if the last trade price is higher than the most recent trade at a different price.<sup>89</sup>

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.<sup>90</sup> 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 at a low price.

Rule 202T allowed the SEC to implement a “pilot program to examine the efficacy” of the short sale restrictions.<sup>91</sup> 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.<sup>92</sup> So long as it is effectively random which 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.<sup>93</sup>

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

---

<sup>89</sup> 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*”).

<sup>90</sup> *Id.* at 6-8.

<sup>91</sup> *Id.* at 4.

<sup>92</sup> Stratified sampling occurs because:

in any randomized trial it is desirable that the comparison groups should be as similar as possible as regards participant characteristics that might influence the response to the intervention.

Stratified randomization is used to ensure that equal numbers of participants with a characteristic thought to affect prognosis or response to the intervention will be allocated to each comparison group . . . . Stratified randomization is performed either by performing separate randomization (often using random permuted blocks) for each strata, or by using minimization.

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.

<sup>93</sup> *SEC Report*, *supra* note 89, at 4.

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.”<sup>94</sup> 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.<sup>95</sup>

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.<sup>96</sup> In addition, the report explains the methodological tools applied to examine the impact of the short sales restrictions on various outcomes.<sup>97</sup>

Finally, the report provides a detailed examination of 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.<sup>98</sup> 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).<sup>99</sup> 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.<sup>100</sup> 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

---

<sup>94</sup> *Id.* at 4.

<sup>95</sup> *Id.* at 6-10.

<sup>96</sup> *Id.* at 22-28.

<sup>97</sup> *Id.* at 28-34.

<sup>98</sup> *Id.* at 34-51.

<sup>99</sup> *Id.* at 51-56; *see also id.* tbls. 3 & 6.

<sup>100</sup> *See id.* Appendix A.

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 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.

## 2. 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.<sup>101</sup>

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.<sup>102</sup> 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. 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

---

<sup>101</sup> The internal controls requirements obligated companies to set up elaborate mechanisms for detecting malfeasance within the company or disclose the absence of such controls.

<sup>102</sup> 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 Xiyang Zhang, *Economic Consequences of Sarbanes-Oxley Act of 2002*, 44 J. ACCT. & ECON. 74 (2007).

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 (xxx find others). 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 subject 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's 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.<sup>103</sup> Instead, the experimental repeal should be applied for an extended period—up to several years.<sup>104</sup>

Just as in the short sale experiment, the unit of observation for an experimental Sarbanes-Oxley repeal should be the publicly traded company. Sarbanes-Oxley's requirements apply to publicly traded corporations, making the choice of unit of observation relatively straightforward. If, however, as discussed above, SOX repeal is likely to produce substantial competitive advantages for untreated firms (i.e., those still subject to SOX requirements),<sup>105</sup> 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

---

<sup>103</sup> 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.

<sup>104</sup> 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.

<sup>105</sup> 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.

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.<sup>106</sup>

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 “subject” SOX-repeal group and others to a “control” group with SOX remaining in place.<sup>107</sup> 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 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.<sup>108</sup>

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.

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

---

<sup>106</sup> 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.

<sup>107</sup> 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.

<sup>108</sup> *Supra* note 119.

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.

Two important alternative measures of SOX's efficacy are stock market value and listing/delisting decisions. 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 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.

Companies are more likely to list on public markets if the benefits (access to capital, liquidity, etc.) outweigh the costs—public filings, risk of lawsuits, etc. SOX alters both the benefits and the costs of listing. SOX should be randomly assigned to all companies applying to list on U.S. markets, as well as companies already listed. If SOX's benefits outweigh its costs, then companies (both foreign and domestic) that apply to list on the stock market and are assigned to SOX will be more likely to complete the listing process than companies that are not assigned to SOX when they are assigned to SOX. By observing companies voting with their feet, experimenters can gain another objective measure of corporate impressions of SOX's efficacy. Similarly, if currently listed companies view SOX as not worth the trouble, then companies that are randomly assigned to SOX will be more likely to delist<sup>109</sup> than companies that are not subject to SOX. Observing whether these companies continue to pursue the listing, experimenters can get further evidence about SOX's efficacy.<sup>110</sup>

At the end of the experimental period, the outcomes for the SOX and non-SOX groups should be examined. Given the intense academic interest in SOX, there will likely be a large number of studies of the experiment, allowing for greater confidence that any conclusions about the experiment will be subject to extensive vetting. An experimental repeal of SOX for some companies but not others is therefore likely to provide important new evidence about the true effects of this important but extremely controversial policy.

Again, there is nothing unique about SOX. 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. Indeed, the value of experiments is by no means

---

<sup>109</sup> There are several methods of delisting, including going private, choosing to list on non-U.S. markets, or being acquired by another company. Note that to prevent companies assigned to SOX from avoiding SOX by merging with shell companies that are not in the SOX group, there should be a rule that the SOX status of the larger company will apply whenever two companies with different SOX statuses merge.

<sup>110</sup> Delisting decisions may be driven by factors other than value, of course. For example, managers may dislike SOX's requirements even though they enhance value. Powerful managers may therefore try to delist to avoid SOX even if SOX raises value. As a result, delisting decisions must be compared with listing decisions (which suffer from less of a principal/agent problem) and market value impacts (among others) to gain a true measure of the net effects of SOX.

confined to securities law. The next section, for example, examines the value of experiments for resolving tax policy debates.

## **B. Tax Law Experiments**

Few topics in public policy are as hotly debated as tax policy. In particular, economists debate the impact of different tax rates on incentives to work. Economists affiliated with the Republican Party 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.<sup>111</sup> Others argue that hours and entrepreneurship are not particularly sensitive to relatively small changes in marginal tax rates, meaning that government revenues will fall nearly 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).<sup>112</sup>

Tax rates change frequently, so there is ample variation with which to study how the change in tax rates impacts labor supply and entrepreneurship.<sup>113</sup> 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.<sup>114</sup> 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.<sup>115</sup>

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

---

<sup>111</sup> If a change in tax rates has no impact on behavior, then the revenue loss can be estimated by the decrease in the tax rate. Most economists, however, think that a change in the tax rate has some effect on the supply of labor and entrepreneurship. Some economists even claim that lowering tax rates can increase revenue, but this claim is discredited. N. Gregory Mankiw, *The Optimal Collection of Seigniorage Theory and Evidence*, 20 J. MONETARY ECON. 327 (2004).

<sup>112</sup> 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.

<sup>113</sup> 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).

<sup>114</sup> 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).

<sup>115</sup> 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.

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.<sup>116</sup> The critical outcomes of interest in the tax debate is the impact of tax rates on labor supply and entrepreneurship. These decisions are made at the individual or household level, meaning that individuals or households are the appropriate units of observation.<sup>117</sup>

Imposing differential mandatory tax rates on similarly situated individuals will undoubtedly be controversial. There are several means of mitigating this controversy to allow the experiment to take place. 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 effect<sup>118</sup>—then the experiment fails in its aim to provide conclusive evidence about the impact of marginal tax rates on labor supply and entrepreneurship.

Alternatively, the government can ask for volunteers for a tax experiment that randomly assigns individuals to a subject group with lower taxes or a control group with tax rates identical to the status quo. At least some individuals will volunteer because the expected tax rate from participating in the experiment is lower than the tax rate from non-participation. The differential tax treatment received by these volunteers should be less controversial than a mandatory increase because the individuals subject to the experiment will have consented to the unequal treatment. Volunteers, of course, raise the heterogeneous treatment effect problem discussed in Section IV.B. If volunteers have different responses to marginal tax rates than the average individual, the experiment will not provide an unbiased estimate of the effect of changes in marginal tax rates for the whole population. For example, volunteers may be more aware of and sensitive to tax rates than the average individual, making the labor supply response to different tax rates

---

<sup>116</sup> Note that by varying the unit of randomization between the individual and the household, policymakers can get a sense of the true effect of the “marriage penalty,” 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.

<sup>117</sup> 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.

<sup>118</sup> 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).

higher for volunteers than for the average individual. The experimenter must weigh the benefits of volunteerism against this risk of biased treatment effect estimates.

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.

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 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).

A randomized experiment assigning different marginal tax rates to different individuals would provide compelling evidence regarding one of the most salient contemporary public policy debates—the extent to which lowering taxes changes individual behavior. This taxation experiment provides yet another example of the generality and utility of systematic randomized experimentation of public policy.

### ***C. An ENDA Experiment***

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. Randomized experiments could provide powerful evidence of alternative criminal laws. For example, if different jurisdictions were randomly assigned to have or not have a death penalty for ten years, we could look to see whether there were appreciable differences in the death penalty.

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 orientation.<sup>119</sup> The Employment Non-

---

<sup>119</sup> 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](http://www.hrc.org/documents/employment_laws_and_policies.pdf) (California (1992, 2003), Colorado (2007), Connecticut (1991), District of Columbia (1977, 2006), Hawaii (1991), Illinois (2006), Iowa (2007), Maine (2005), Maryland (2001), Massachusetts (1989), Minnesota (1993), Nevada (1999), New Hampshire (1998), New Jersey (1992, 2007), New Mexico (2003), New York (2003), Oregon (2008), Rhode Island (1995, 2001), Vermont (1991, 2007), Washington (2006), and Wisconsin (1982).)

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. But ENDA still faces substantial opposition (including President George W. Bush and Republican Presidential Nominee John McCain). Even though polls suggest that an overwhelming majority of Americans oppose employment discrimination on the basis of sexual orientation,<sup>120</sup> opponents argue that ENDA would impose substantial litigation and other compliance costs that would be visited on private employers. For example, a “Statement of Administration Policy” issued by President Bush’s Office of Management and Budget concludes that “his senior advisors would recommend that he veto the bill” in part because:

The bill turns on imprecise and subjective terms that would make interpretation, compliance, and enforcement extremely difficult. For instance, the bill establishes liability for acting on “perceived” sexual orientation, or “association” with individuals of a particular sexual orientation. If passed, [ENDA] is virtually certain to encourage burdensome litigation beyond the cases that the bill is intended to reach.<sup>121</sup>

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.<sup>122</sup> 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.<sup>123</sup> 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.<sup>124</sup>

---

Approximately half of non-farm employees are protected by these laws. See SEAN CAHILL, NAT’L GAY AND LESBIAN TASKFORCE, *THE GLASS NEARLY HALF FULL: 47% OF U.S. POPULATION LIVES IN JURISDICTION WITH SEXUAL ORIENTATION NONDISCRIMINATION LAW 1* (2005), available at <http://www.thetaskforce.org/downloads/GlassHalfFull.pdf>.

<sup>120</sup> 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).

<sup>121</sup> The statement also raised First (free exercise of religion) and Eleventh (state immunity) Amendment concerns. Statement of Administration Policy, H.R. 3685 (Oct. 23, 2007), available at <http://www.whitehouse.gov/omb/legislative/sap/110-1/hr3685sap-r.pdf>. When challenged by Al Gore in a 2000 presidential debate on whether he would support ENDA, candidate Bush responded: “I don’t think they ought to have special rights, but I think they ought to have the same rights.” When asked to define “special rights,” Bush said, “Well, it’d be if they’re given special protective status.” *Newshour with Jim Lehrer, Second Presidential Debate* (PBS television broadcast Oct. 11, 2000) (transcript available at <http://www.pbs.org/newshour/bb/election/2000debates/2ndebate3.html>).

<sup>122</sup> 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).

<sup>123</sup> 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).

<sup>124</sup> *Id.*

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 whether the opponents' objections are well founded. We tend to think that Congress faces an all-or-nothing decision—to prohibit or not to prohibit employment discrimination. But as we've seen before, an intermediate approach would be to pass ENDA for a randomly assigned subset of employers. 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. To begin, an RTC would need to determine the unit of randomization. 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. If the opponents are right, firms forced to comply may be put at a competitive disadvantage and it might be appropriate to randomize at the industry level, so that covered firms would not have to compete against uncovered firms. But 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. (Indeed, 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.<sup>125</sup>

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 employees<sup>126</sup>—that it would be possible to perform a powerful test that assigns as few as 10% 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.

The content of “treatment” law itself is of vital importance. The language of ENDA represents in many ways an irreducible civil rights minimum. ENDA has been

---

<sup>125</sup> 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 the number of observations.

<sup>126</sup> U.S. CENSUS BUREAU, CENSUS, tbl. 735, *available at* <http://www.census.gov/compendia/statab/tables/08s0735.pdf>.

carefully crafted solely to prohibit disparate treatment. It does not allow disparate impact claims or claims for health benefits by unmarried domestic partners. And it expressly does not require affirmative action.<sup>127</sup> But there is one dimension of coverage on which proponents of the bill sharply disagree, and that is whether the bill should also prohibit discrimination on the basis of “gender identity” as well as sexual orientation. In April 2007 a version of ENDA (sometimes referred to as “GENDA”) which includes “gender identity” in the discrimination coverage was introduced into the House of Representatives.<sup>128</sup> However, in September 2007, Representative Barney Frank and others sponsored the more restrictive version of ENDA in the face of opposition from more than three hundred LGBT organizations.<sup>129</sup> In November, the more restrictive version passed the House, but to date the Senate has not considered either version of the statute.

Some opponents of GENDA are particularly concerned about the costs of compliance and raise a parade of horrors, suggesting the prohibition on gender identity discrimination would interfere with employers’ reasonable dress code standards. To our minds, the GENDA/ENDA dispute is also susceptible to randomized testing. One could divide firms into three groups, with at least 5% of firms assigned to comply with GENDA and 5% being assigned to comply with ENDA. A three-way randomization could determine the marginal impact of “gender identity” on a host of economic variables.<sup>130</sup>

A more libertarian version of the test would merely assign different ENDA defaults to different firms. Federal law currently allows employers to intentionally discriminate on the basis of employee sexual orientation. But this employer freedom to discriminate is nothing more than a default. There is nothing to stop employers from opting into ENDA by private contract and giving their employees and applicants virtually identical rights, including private rights of action, as they would have if ENDA passed. Indeed, Jennifer Brown and Ian Ayres have created a contractual mechanism where any employer with just a few clicks at [www.fairemploymentmark.org](http://www.fairemploymentmark.org) can do just that.<sup>131</sup> 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.<sup>132</sup>

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

---

<sup>127</sup> Ayres & Brown, *supra* note 123.

<sup>128</sup> Employment Non-Discrimination Act of 2007, H.R. 2015, 110th Cong. (2007).

<sup>129</sup> Employment Non-Discrimination Act of 2007, H.R. 3685, 110th Cong. (2007); *see also* LAMBDA LEGAL, LAMBDA LEGAL’S ANALYSIS OF H.B. 3685 (2007), *available at* [http://data.lambdalegal.org/pdf/enda\\_llanalysis\\_20071016.pdf](http://data.lambdalegal.org/pdf/enda_llanalysis_20071016.pdf). HRC (the Human Rights Campaign) was one of the only major LGBT organizations not to oppose the passage of the more restrictive coverage.

<sup>130</sup> Or, if a concern over the potential litigation costs induced by coverage of “perceived” orientation is truly a stumbling block for some lawmakers, it would be possible to test for the marginal impact of including or excluding this group in the Act’s coverage.

<sup>131</sup> The fair employment license falls short of ENDA protections in a few dimensions. *See* Ayres & Brown, *supra* note 123, 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).

<sup>132</sup> Ian Ayres, *Menus Matter*, 73 U. CHI. L. REV. 3 (2006).

defaults and menus matter.<sup>133</sup> Instead of running an RTC 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.<sup>134</sup>

## 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 policies. It remains the case, however, that the vast majority of policy changes are enacted without the benefit of randomized evaluations. This Article seeks to systematize and expand the use of randomized experiments of law and policy. Instead of agencies or legislatures implementing or (more often) not implementing randomized experiments on an ad hoc basis, we propose that government conduct randomized experiments of all new policies, with few exceptions. We argue that a randomization impact statement (RIS) should be required for all new regulations. In an RIS, agencies would be required to disclose experimental estimates for the efficacy of a policy or explain the absence of such experiments. The agencies would then have to explain why the experimental evidence favors a particular policy. We believe that an RIS can improve agency decision-making in much the same way as cost benefit analyses and environmental impact statements rationalize public policymaking along other dimensions. While legislatures should not be required to describe experimental evidence in favor of or against various policies, a norm in favor of experimental evidence rather than other types of evidence would significantly alter the empirical background of policy debates.

---

<sup>133</sup> Yair Listokin, *What Do Corporate Default Rules and Menus Do? An Empirical Examination* (Yale Law School, Working Paper No. 335, 2005).

<sup>134</sup> 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).