Rigorous Policy Pilots: Experimentation in the Administration of the Law

Colleen V. Chien[*]

ABSTRACT: Rigorous tests are being used every day to develop effective medical treatments, drive consumer engagement, and, more generally, discover what works. But so far, rigorous policy piloting—temporarily introducing a change in law or policy in order to learn from it using well-designed and well-implemented methods—has not been used widely because of the perception that policy experimentation is unfair and possibly illegal, difficult, and rare. This Essay draws upon case law and agency practice to show that, to the contrary, rigorous policy pilots are presumptively legal, feasible, and increasingly common, proceeding in several steps. First, it finds that many kinds of pilots, including those that vary internal agency processes, or which are opt-in are unlikely to be controversial. But a review of relevant cases suggests that courts are likely to uphold even pilots that treat like members of a population differently, including through randomization, when they advance learning. Further, it finds experimentation, by itself, to be unlikely to create special procedural or substantive hurdles. Second, it finds that agencies are engaging in a range of rigorous piloting activities to fill informational gaps in policy- and law-making, some of which simulate and others which effect policy variation on a temporary basis, and that

[*] Justin D'Atri Visiting Professor of Business Law, Columbia Law School; Professor, Santa Clara University School of Law; 2013–2015 White House Senior Advisor, Innovation and Intellectual Property. A companion online appendix to this Essay with policy experiments the USPTO could try can be found at Colleen V. Chien, *Rigorous Policy Pilots the USPTO Could Try*, 104 IOWA L. REV. ONLINE 1 (forthcoming 2019). This Essay was the subject of a companion workshop, *Rigorous Policy Pilots*, co-sponsored by the Administrative Conference of the United States, Penn Program on Regulation, Columbia Center for Constitutional Governance, Partnership for Public Service, and Santa Clara High Tech Law Institute held in 2019. Materials from that workshop, a list of federal agency experiments and evaluation components, and relevant readings and resources are posted to which are available at: https://www.law.upenn.edu/institutes/ppr/policypilots. I thank Jenna Clark, Jonathan Liu, Lauryn Younge, Jiun-Ying Wu, Arti Rai, Bernard Chao, David Schwartz, Cary Cognialese, Todd Rubin, Reeve Bull, Dan Correa, Lindsay Laferriere, Michael Abramowicz, Heidi Johnson, Jonah Probell, Chris Walker, Bert Huang, Olatunde Johnson, Jessica Bulman-Pozen, Richard Briffault, Eric Talley, Stuart Graham, Rose Cuillon-Villasor, Suzanne Kim, David Noll, and audiences at the *Iowa Law Review* Symposium on Administering Patent Law, MIT Media Lab, and Unified Patents Conference for their helpful input to this project.

developments such as the growth of open data are making such forms of information gathering easier. It draws from agency experience to develop a framework for proposing a policy pilot and identify steps that would further support the use of rigorous pilots. A companion online appendix applies this framework to propose several rigorous pilots that the United States Patent and Trademark Office ("USPTO"), building on its already strong tradition of piloting, to evolve its own policies and practices with respect to patent quality (through the robust vetting of applications in view of non-patent literature and team/time examination on demand) and inclusion in innovation (through automated error correction and addressing gender bias in examination).

I. INTRODUCTION

> "The country needs and, unless I mistake its temper, the country demands bold, persistent experimentation. It is common sense to take a method and try it: If it fails, admit it frankly and try another. But above all, try something."[1]
>
> - Franklin Delano Roosevelt, Address at Oglethorpe University, May 22, 1932

In the fall of 1960, the Food and Drug Administration ("FDA") received an application for a drug for treating morning sickness. Already approved and sold in dozens of countries,[2] "Kevadon" was well-positioned for approval and entry into the United States. But its reviewer, a Canadian-born physician named Frances Kelsey, wasn't convinced. Over several months and rounds of document exchanges, she continued to find evidence of the drug's safety and effectiveness lacking. The manufacturer mounted a high-pressure campaign and complained about Kelsey, whom they called a "petty bureaucrat" to her boss.[3]

The drug—better known by its generic name, thalidomide, and never approved for morning sickness in the United States—was later revealed to be the source of numerous horrific birth defects.[4] Kelsey was lionized and bestowed with honors.[5] But while Kelsey's fame is well-deserved, the story has a darker legacy. As Vincent DeVita, former head of the National Cancer Institute, describes in *The Death of Cancer*, this episode "sent the message to those who worked at the FDA that the way to do right by people was to say no."[6] The subsequently enacted Drug Regulation Act put drug discovery out of the reach of all but the largest companies[7] and bred the risk-aversion for which the agency is known,[8] according to critics. One way the agency stays out of trouble is to keep risky medicines off the market, a good thing. But there

---

1. Franklin Delano Roosevelt, Address at Oglethorpe University (May 22, 1932), *available at* http://www.fdrlibrary.marist.edu/_resources/images/msf/msf00486.

2. *Frances Oldham Kelsey: Medical Reviewer Famous for Averting a Public Health Tragedy*, U.S. FOOD & DRUG ADMIN., https://www.fda.gov/aboutfda/history/virtualhistory/historyexhibits/ucm345094.htm (content current as of Feb. 1, 2018).

3. Robert D. McFadden, *Frances Oldham Kelsey, Who Saved U.S. Babies from Thalidomide, Dies at 101*, N.Y. TIMES (Aug. 7, 2015), https://www.nytimes.com/2015/08/08/science/frances-oldham-kelsey-fda-doctor-who-exposed-danger-of-thalidomide-dies-at-101.html.

4. VINCENT DEVITA & ELIZABETH DEVITA-RAEBURN, THE DEATH OF CANCER 195 (2015).

5. Including as the second female recipient of the President's Award for Distinguished Federal Civilian Service. *Dr. Frances Kathleen Oldham Kelsey*, NAT'L INST. HEALTH, https://cfmedicine.nlm.nih.gov/physicians/biography_182.html (last updated June 3, 2015).

6. DEVITA & DEVITA-RAEBURN, *supra* note 4, at 196.

7. *See generally* Hans-Georg Eichler et al., *The Risks of Risk Aversion in Drug Regulation*, 12 NATURE REVIEWS DRUG DISCOVERY 907 (2013) (including among the adverse effects of excessive regulatory risk aversion: slow approval times, the denial of useful medications to patients, and the inability of small companies to compete in the marketplace).

8. *See id.* at 908–11.

are "no awards for getting a good drug quickly into the public domain."[9] In fact after a prolonged delay, thalidomide returned to the market as an effective treatment for leprosy, blood cancer, and multiple myeloma.[10]

In the government, fear of making a Type 1 or "false positive" error —taking a mistaken step—leads to an overabundance of Type 2 or "false negative" errors—failing to take a productive step. At root, however, is the problem of "what [the] government doesn't know," regarding what changes to law or policy to consider taking, what the impact of such changes might be, and which among several potential changes is most likely to achieve a policy goal,[11] all topics about which there may be many opinions, but little relevant evidence. Uncertainty about whether a law or policy change will achieve its intended purpose, fear of making a mistake in the face of this uncertainty, and institutional inertia all contribute to status quo bias—whether in favor of preserving regulations that don't work or failing to adopt new policies that do.[12]

One way to address both the knowledge gap and risk aversion in policy development is by introducing a temporary change to law or policy for the purpose of learning from it, through a "policy pilot." Implemented with rigor, through the application of, "well-designed and well-implemented methods tailored to the questions being asked"[13]—rigorous policy pilots are a generative yet under-used tool for addressing informational deficits that stand in the way of developing effective law and policy.

The range of open questions in law and policy that can be addressed by pilots are wide-ranging. They include, for example, the question of how giving a universal basic income ("UBI") to individuals impacts their well-being and employment. In 2017, the Finnish government began tracking the outcomes of 7,000 unemployed citizens, about a third to which it gave $600 a month,

---

9. DEVITA & DEVITA-RAEBURN, *supra* note 4, at 196.

10. P.J. Lachmann, *The Penumbra of Thalidomide, the Litigation Culture and the Licensing of Pharmaceuticals,* 105 QJM: INT'L J. MED. 1179, 1179 (2012).

11. See discussion in Cass R. Sunstein, *Financial Regulation and Cost-Benefit Analysis*, 124 YALE L.J. FORUM 263, 263–64 (2015) (describing the "knowledge problem" in law- and policy-making that "public officials face in attempting to obtain relevant information, much of which is widely dispersed in society").

12. William W. Buzbee, *Interaction's Promise: Pre-emption Policy Shifts, Risk Regulation, and Experimentalism Lessons*, 57 EMORY L.J. 145, 156 (2007) (describing regulatory status quo bias and stickiness, and risk-aversion among regulators).

13. H.R. REP. NO. 115-411, at 2 (2017) (designating, in addition to "rigor: . . . [the use of] well-designed and well-implemented methods tailored to the question being asked," privacy, transparency, humility, and capacity as a guiding principle of evidence-based policymaking); Foundations for Evidence-Based Policy Making Act of 2018, Pub. L. No. 115-435 [hereinafter *2018 Evidence Act*]; *see also* No Child Left Behind Act of 2001, Pub. Law. No. 107-110, 15 Stat. 1525 (codified at 20 U.S.C. § 6368(6)(A) (2006)) (defining scientifically based research as "research that applies *rigorous*, systematic, and objective procedures to obtain valid knowledge relevant to . . . [education activities and programs]" (emphasis added)).

to address this question;[14] several U.S. jurisdictions are planning their own tests.[15] Also unknown is how the administration of patent law and issuance of quality patents best advance the Constitutional goal of "promot[ing] the progress of science and useful arts."[16] Over the last decade, the United States Patent and Trademark Office ("USPTO") has run numerous pilots to address this question.[17] As the D.C. Circuit has said, "there are some situations in which, 'a month of experience will be worth a year of hearings.'"[18] While experiments or pilots can address a variety of questions pertaining, e.g., to a policy's feasibility or stakeholder reactions, when the question concerns whether the policy has caused an observed outcome, a controlled trial that provides the treatment to one part of population, withholds it from another, and then compares the difference in outcomes provides the best insight. Such rigorous experiments can be designed, in turn, with or without randomization, and to vary the rules or laws that apply or to approximate such changes. A pilot that limits to whom a rule or policy applies constrains the risk but not the learning. As Justice Brandeis famously dissented, "a single courageous state may . . . serve as a laboratory; and try novel social and economic experiments *without risk to the rest of the country*."[19]

Within today's highly partisan political landscape, the need for quality evidence about the performance of law and policy presents a rare opportunity for agreement, on the methods if not the values or agendas advanced by evidence. Calls for "[s]marter use of data and evidence . . . to orient decisions and accountability around service and results"[20] and "an aggressive management agenda . . . that delivers smarter, more innovative, and more accountable government for citizens . . . . by applying . . . evidence about what

14. OLLI KANGAS ET AL., THE BASIC INCOME EXPERIMENT 2017–2018 IN FINLAND, PRELIMINARY RESULTS, 29–30 (2019), http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/161361/Report_The%20Basic%20Income%20Experiment%2020172018%20in%20Finland.pdf?sequence=1&isAllowed=y (finding that well-being increased, but that employment did not, as a result of the treatment).

15. For a list of UBI pilots, see Erin Winick, *Universal Basic Income Had a Rough 2018*, MIT TECH. REV. (Dec. 27, 2018), https://www.technologyreview.com/s/612640/universal-basic-income-had-a-rough-2018 (describing ongoing pilots in Alaska, New York, California, and Louisiana among others).

16. U.S. CONST. art. I, § 8, cl. 8.

17. *See* Colleen V. Chien, *Rigorous Policy Pilots the USPTO Could Try*, 104 IOWA L. REV. ONLINE 1 (forthcoming 2019).

18. Md. People's Counsel v. Fed. Energy Regulatory Comm'n, 761 F.2d 768, 779 (D.C. Cir. 1985) (quoting Am. Airlines, Inc. v. Civil Aeronautics Bd., 359 F.2d 624, 633 (D.C. Cir. 1966)).

19. New State Ice Co. v. Liebmann, 285 U.S. 262, 311 (1932) (Brandeis, J., dissenting) (emphasis added).

20. THE PRESIDENT'S MGMT. COUNCIL & THE EXEC. OFFICE OF THE PRESIDENT, PRESIDENT'S MANAGEMENT AGENDA 4 (2018), https://www.whitehouse.gov/wp-content/uploads/2018/03/The-Presidents-Management-Agenda.pdf.

works"[21] have appeared in President Trump's and President Obama's management agendas, respectively. The passage of the Foundations for Evidence-Based Policy Making Act of 2018 ("2018 Evidence Act"), besides showing that Democrats and Republicans actually can agree on something, will consolidate and coordinate statistical expertise for assessment among rank and file bureaucrats across the federal government.[22]

But while "adequate and well controlled studies" have long been required in certain contexts,[23] experiments with laws and rules have a history of being constitutionally and legally suspect. As the Supreme Court stated in *Truax v. Corrigan*, "the Constitution was intended—its very purpose was—to prevent experimentation with the fundamental rights of the individual."[24]

Rigorous pilots that provide a benefit to some, but not other, members of a population would appear to offend basic notions of equal protection.[25] Imposing a regulatory burden on a subset of regulated entities likewise sounds in unfairness and risks running afoul of the Constitution's due process guarantees. Applying a rule randomly to one set of entities but not another would appear to some to present just the kind of "arbitrary, capricious" agency action judicial review was designed to set aside.[26]

How rigorous testing can support evolution of the law is also not widely understood. By operating within existing legal frameworks, experiments can obscure more salient questions, including whether or not the framework is the right one in the first place. Rigorous experimentation is less flexible and adaptable than "democratic experimentalism" with which it shares some features.[27] While Congress may feel empowered to take bolder steps when laws are framed as reversible and temporary, requirements of rigor can constrain agencies, feeding anti-regulatory impulses. Experimental evidence can be weaponized and politicized in the same way non-experimental evidence can.[28]

---

21. Memorandum from Sylvia Burwell et al., Dir., OMB to the Heads of Exec. Dep'ts & Agencies (July 26, 2013) ("President [Obama] recently asked his Cabinet to carry out an aggressive management agenda . . . that delivers a smarter, more innovative, and more accountable government for citizens.").

22. Evidence Act of 2018, Pub. L. No. 115-435, §§ 314–315, H.R. 4174 (2019).

23. Namely in drug approval contexts as provided for by the Drug Efficacy Amendment of 1962 and enshrined in 21 C.F.R. § 314.126 (2018).

24. Traux v. Corrigan, 257 U.S. 312, 338 (1921).

25. *See infra* Part II.

26. Pursuant to the Administrative Procedure Act, 5 U.S.C. § 706(2)(a) (2012).

27. Charles Sabel & William Simon, *Democratic Experimentalism*, *in* SEARCHING FOR CONTEMPORARY LEGAL THOUGHT 477–98 (Justin Desautels-Stein & Christopher Tomlins eds., 2017), *available at* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2983932 (finding parallels in the features of randomized control trials and experimentalist organizations including rolling rules, root-cause analysis, peer review, and performance measurement).

28. For an account of the multiple ways in which evidence can be improperly constructed, manipulated, politicized, and misrepresented in policymaking, replete with specific agency examples, as well as a description of practices for guarding against these outcomes through, e.g., pre-registration of evaluation protocols, transparency of research reports, and ex ante decision

And yet, when compared to doing nothing or engaging in unreflective action, the benefits of rigorous experimentation become clearer: Rigorous pilots can prime legislative and policy-making efforts by underscoring the importance of evidence and prompting a discussion about the metrics that matter. Rigorous pilots can be used to test legal theories, and point out legal infirmities, provided that is the motivation. Further, among information sources, rigorous pilots are distinct in their ability to generate experiential feedback on the operation of law and policy.

But while rigorous policy pilots can thus be powerful tools, so far, they have been underused tools for informing the development of law and policy. In part this is due to a perception that certain types of rigorous pilots are possibly illegal. Citing a litany of difficulties, a recent recommendation of the Administrative Conference of the United States[29] concluded that, "legal, policy, and ethical challenges . . . may mean that regulatory agencies should use randomized study methods only under limited circumstances."[30] But rigorous pilots have also been underused because of practical obstacles to designing and carrying them out, and a lack of awareness of the value they can add.

This Essay attempts to chip away at perceived legal, institutional, and informational barriers to the use of rigorous policy pilots. While informed by the growing academic interest in experiments,[31] it primarily draws upon perhaps the single most credible source for addressing the legality, importance, and mechanics of structured pilots by the federal government: the federal government itself. Based upon caselaw, legislative, and agency accounts of government experimentation, it finds that from being illegal, hard, and rare, rigorous government pilots are presumptively legal, feasible, increasingly common, and in many cases, worthwhile, and proposes a framework for proposing rigorous policy pilots based on these experiences. A companion online appendix, using the patent system as a case study, applies

commitments, see Kristen Underhill, *Broken Experimentation, Sham Evidence-Based Policy*, YALE L. & POL'Y REV. (forthcoming 2019) (draft on file with the author).

29. An independent government agency that offers guidance and best practices to agencies through, *inter alia*, reports and recommendations, as described at *Guidance to Federal Agencies*, ADMIN. CONF. U.S., https://www.acus.gov/guidance-federal-agencies (last visited June 8, 2019).

30. ADMIN. CONFERENCE OF THE U.S., RECOMMENDATION 2017-6, LEARNING FROM REGULATORY EXPERIENCE 5 (2017), https://www.acus.gov/sites/default/files/documents/Recommendation%202017-6%20%28Learning%20from%20Regulatory%20Experience%29_0.pdf.

31. On the emphasis and possibly, overemphasis, on rigorous experimental work in public economics and international development, see Christopher J. Ruhm, *Shackling the Identification Police?* 1–2 (Nat'l Bureau of Econ. Research, Working Paper No. 25320, 2018), *available at* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3294925 (tracing the growth in studies using experimental or quasi-experimental methods among NBER working papers, development evaluations, and top general and field journals). For studies in law academia, see, e.g., Lisa Larrimore Ouellette, *Patent Experimentalism*, 101 VA. L. REV. 65 (2015) (discussing different ways policies can promote innovation in patent experimentation) and experiments by Schwartz, Buccafusco and others described in Part II.

this framework to suggest rigorous policy pilots that the USPTO could implement to support patent law- and policy-making.

Part II reviews relevant case law and finds that rigorous policy experiments raise no particular legal barriers to agency action. Many pilots will be uncontroversial because they do not require the mandatory, selective application of law or policy to the public. But pilots that do, including through randomization, have by and large been found by courts to pass constitutional muster on the basis that they further learning and other legitimate government objectives. The requirements that the Administrative Procedure Act ("APA") imposes will also vary depending on the design and features of a pilot but should not present special procedural hurdles to rigorous pilots nor elevated substantive hurdles to the use of the evidence they generate.

Having established in Part II that rigorous policy pilots are presumptively legal, Part III draws upon decades of rigorous agency experiments to show that, while not necessarily easy, rigorous policy pilots are both feasible and worthwhile in a variety of contexts. It nests lessons that can be learned from these experiences into a basic framework for proposing a rigorous policy pilot, "MATTER," that comprises the following steps: attending to questions that **m**atter, relevant **a**uthority, the underlying **t**heory of change, **t**esting protocol, **e**vidence and **r**esources.[32] It aspires to audiences of not only policymakers but also those engaged in forming policy recommendations, on the theory that framing new policy ideas as pilots to learn from, not just new policies to adopt, can increase their uptake. Part IV concludes, and an online companion appendix applies this framework to the patent system to suggest several pilots that the Patent Office could try.

II. RIGOROUS POLICY PILOTS ARE PRESUMPTIVELY LEGAL

Every day, companies are developing prototypes and carrying out structured tests to find effective treatments[33] and increase consumer engagement.[34] Why, to date, hasn't rigorous piloting been widely used to develop and guide the evolution of federal law and policy, improving their effectiveness for the American people? Though there are many reasons, the

32. Referable to as "MATTER" (questions that **m**atter, relevant **a**uthority, the underlying **t**heory of change, **t**esting protocol, **e**vidence and **r**esources).

33. E.g., to gain approval from the FDA pursuant to 21 C.F.R. § 314.126 (2018), which requires "adequate and well-controlled studies."

34. *Hearing on What Works / Evaluation Before the Subcomm. on Human Resources of the H. Comm. On Ways & Means*, 113th Cong. 3 (2013) (Statement of Jon Baron, President, Coalition for Evidence-Based Policy) (describing the 13,000 randomized trials of new products/strategies conducted by Google and Microsoft, 80–90% of which have reportedly had no significant effects); Eytan Bakshy, *Big Experiments: Big Data's Friend for Making Decisions,* FACEBOOK (April 3, 2014), https://www.facebook.com/notes/facebook-data-science/big-experiments-big-datas-friend-for-making-decisions/10152160441298859 [http://tinyurl.com/yyjv92h2] (Facebook empiricist describing how the company "run[s] over a thousand experiments each day. While many of these experiments are designed to optimize specific outcomes, others aim to inform long-term design decisions.").

robustness of available evidence concerning the effectiveness of law and policy is not one of them. The fragility of government feedback loops is well-documented and structural, stemming from, *inter alia*, the absence of competition,[35] the anecdotal and fragmentary nature of litigation, short democratic terms and broad Congressional mandates, and the prioritization, in public policy, of the urgent over the important.[36] The dynamics of capture and the non-participation of impacted entities in processes like notice-and-comment and town halls, or what could be called the "turn-out problem" in lawmaking, limit the quality of feedback.[37]

Against this backdrop, the promise of objective, real-world data on the performance of law or policy from a pilot might seem irresistible. However, for the average federal agency contemplating an experiment that will selectively burden or benefit members of the public, Constitutional, APA, and basic fairness considerations loom large.[38] The act of designating treatment and control groups in effect "allow[s] . . . officials to pick and choose only a few to whom they will apply legislation," a practice condemned by Justice Jackson in *Railway Express Agency, Inc. v. New York* as contrary to the Constitution.[39] Random selection can feel "arbitrary-and-capricious" and potentially unlawful. Pilots that induce capital-intensive investments or that specify or result in differential levels of health or safety protection may, worryingly, place one set of regulated entities or members of the public at a disadvantage relative to another.[40]

The case for caution in government experimentation is historically grounded. In the 1970's, revelations about the withholding of effective treatments from African-American men with syphilis in Tuskegee, Alabama, by the U.S. Public Health Service to complete a government experiment set

35. D. James Greiner & Andrea Matthews, *Randomized Control Trials in the United States Legal Profession*, 12 ANN. REV. L. & SOC. SCI. 295, 296 (2016) ("[T]he construction and administration of adjudicatory systems. . . . [represent] arenas [that], unlike those in which legal professionals and judges compete for business, lack the discipline that markets can sometimes impose on inefficient or wasteful practices.").

36. As elaborated with flourish by Cass R. Sunstein, *The Most Knowledgeable Branch,* 164 U. PA. L. REV. 1607, 1608–1611 (describing the "grotesquely distorting prism of litigation," and that, due to the press of time and electoral incentives, members of Congress "might also have no idea what they're talking about," but also cautioning that "happy talk," or the risk of like-minded people coalescing around falsehoods can afflict the executive as well).

37. *See, e.g.*, COLLEEN V. CHIEN & TOM COTTER, REDESIGNING PATENT LAW (forthcoming 2020) (comparing amicus and comment participation in the patent system and finding a high and relatively higher concentration of patent lawyers and owners as commenters to the patent office as compared to amici in patent cases).

38. ADMIN. CONFERENCE OF THE U.S., *supra* note 30, at 5. concluding, based on a report developed in consultation with multiple agencies, that "randomized study methods may present legal, policy, and ethical concerns.").

39. Railway Exp. Agency v. New York, 336 U.S. 106, 112 (1949) (Jackson, J., concurring).

40. *See* ADMIN. CONFERENCE OF THE U.S., *supra* note 30, at 5.

off a chain of events that led eventually to the Common Rule,[41] a set of ethical and legal commitments that govern federally-funded research on human subjects.[42] In their concurrence in *United States v. Stanley*, a case involving the Army's unconsented experimentation to test the impacts of lysergic acid diethylamide ("LSD") on servicemen, Justices Brennan, Marshall, and Stevens rebuked "the Government of the United States [for] treat[ing]thousands of its citizens as though they were laboratory animals."[43] But even less dramatic examples can give pause. Would you want to be the unemployed person that does not receive a universal basic income?[44] As a small business owner, how would you feel about being included in a pilot that subjects your patent application, but not that of others, to intensified scrutiny?[45]

How can desires for fairness and quality information be squared? This Part considers the Constitutional and APA requirements that apply to agency policy pilots. Before doing so, however, it is important to keep in mind that the range of possible pilots an agency can run is as vast and varied as the scope of agency action. What makes a pilot rigorous is that it uses "well-designed and well-implemented methods tailored to the questions being asked."[46] As such, the design of each pilot will depend highly on context and motivation, with implications for its legality. When the relevant question concerns whether a policy change is effective (rather than just feasible or popular, based on its uptake), a rigorous approach will require the identification of a control and treatment group, which can—but does not necessarily—raise equal protection and fairness concerns. Below I describe in simplified terms, several major classes of controlled experiments before discussing their likely treatment under the law (from least to most legally challenged), providing examples from the universal basic income ("UBI") and patent pilots referred to in the introduction to illustrate each concept.

Natural Experiments: In some cases, "control" and "treatment" groups are created organically, by the operation of law or policy. Say that a country's UBI stipend is distributed only to individuals born after January 1, 1970. Individuals just before the eligibility date can be put into a "control" group

41. 45 C.F.R. § 46(A)–(D) (2009). The relevant history is recounted at Jacob Metcalf, *Big Data Analytics and Revision of the Common Rule*, 59 COMM. ACM 31, 31–33(2016).

42. *See Federal Policy for the Protection of Human Subjects ('Common Rule')*, HHS.GOV, https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html (last visited June 8, 2019) (listing the departments and agencies under the Common Rule).

43. United States v. Stanley, 483 U.S. 669, 686 (1987) (Brennan, J., dissenting).

44. As a non-selected participant in one of the numerous universal basic income ("UBI") trials described in Chien, *supra* note 17.

45. The USPTO did a version of this when it did a random audit of already issued trademarks as part of a pilot program to test the integrity of the trademark registry. Out of the 500 trademark registrations subject to the pilot, 51% were unable to verify the previously claimed use, resulting in cancellation or deletion of parts of the registration. *See generally* USPTO, POST REGISTRATION PROOF OF USE PILOT FINAL REPORT (2015) (describing the results of the pilot program).

46. H.R. REP. NO. 115-411, at 2 (2017).

while individuals just after it can be put into a "treatment" group. Differences in their outcomes can be attributed to the treatment. To take an example from the patent system, patent rights are territorial, and the same patent application is often filed in multiple patent offices by applicants seeking coverage in different countries. This creates a natural experiment involving "identical twin" applications, by which the impact of different forms of examination on the same application can be compared.

"Lab" or Simulated Experiments: Another way to carry out controlled policy pilots is through "lab" or simulated experiments. To test how best to distribute a UBI, for example, one could imagine an agency contracting with a lab to run an experiment, as the Consumer Protection Financial Bureau ("CFPB") has done.[47] The pilot could, for example, vary the way that the UBI is distributed, e.g., through a credit card with all the money up front, or through a regular cash disbursement. Any differences in outcomes pertaining, for example, to the well-being and employability of the participants could then be observed.[48] Though doing so wouldn't generate the same information as a full-fledged field pilot, such a pilot could provide suggestive evidence of the impact of the distribution method, if any. Of relevance to the patent system, laboratory methods have been used to test the impact of changes in the standard of proof required to invalidate a patent[49] as well as to test the impact on innovation of patent versus copyright-like incentives.[50] The USPTO has also created pilots that applicant can opt into.[51] Like lab experiments, voluntary programs simulate but differ in important ways from mandatory programs.

Experimenting with Internal Agency Procedures: "Field" experiments vary the experiences of regulated entities. One way of doing so is by experimenting with internal agency procedures. In the case of UBI, for example, one could imagine trying different ways of disbursing stipends to individuals, e.g., through a high- or low-touch process. The outcomes of individuals that receive money one way could then be compared with the outcomes of individuals that receive it another way. In a similar vein, how the Patent Office applies the resource of patent examination can and has been varied across

47. Jeffrey Carpenter et al., *Choice Architecture to Improve Financial Decision Making* 4 (CFPB Office of Research Working Paper Series, Working Paper No. 2017-04, 2017) ("[W]e run a controlled laboratory experiment to evaluate the consequences of the policy intervention before it takes place.").

48. *See id.*; *see also* Heidi Johnson & Jesse Leary, *Policy Watch: Research Priorities on Disclosure at the Consumer Financial Protection Bureau*, 36 J. PUB. POL'Y & MARKETING 184, 187–89 (2017).

49. *See generally* David L. Schwartz & Christopher B. Seaman, *Standards of Proof in Civil Litigation: An Experiment from Patent Law*, 26 HARV. J.L. & TECH. 429 (2013).

50. Christopher Buccafusco et al., *Experimental Tests of Intellectual Property Laws' Creativity Thresholds*, 92 TEX. L. REV. 1921, 1978 (2014).

51. *See* USPTO glossary pilot, described *infra* at Part III.

applications,[52] with certain applications getting more scrutiny than others. Rigorously comparing the outcomes of otherwise equivalent applications can provide suggestive evidence the impact of the extra scrutiny.

Experimenting with the Laws and Rules that Apply to Private Parties: Another way to experiment is by varying the rules that apply, not to internal agency procedures, but to the behavior of private citizens. For example, some citizens in the Finnish UBI pilots were entitled to enroll in the partial basic income program while others were not. In related work, I propose a patent pilot program in which certain applicants are entitled to waive or defer the patent eligibility requirement, and some are not.[53] Comparing the outcomes of "treated" applications with the outcomes of "untreated" applications can generate information about the operation of this legal standard, in view of the other standards of patentability. An agency example comes from the Office of the Comptroller of the Currency ("OCC") which, from 1999–2001 extended special lending limits to 277 community banks as part of a pilot program and compared their outcomes with companies not receiving the treatment.[54]

It is worth pausing here to note that of these four classes of pilots only the last two involve the mandatory application of different procedures or rules to similarly situated entities. When a pilot is implemented as a temporary program that applies to all, equal protection claims are avoided.[55] Differences in treatment that follow the decision to opt into a pilot are the result of free choice, not government discrimination, the Supreme Court has said.[56] These sorts of pilots do not raise the same fairness considerations that the third and fourth categories do and are likely to be less controversial. Even if they do not *vary* law or policy, they can still be useful for *informing* law or policy.

The legality of pilots of the third and fourth types—which vary internal agency processes and the rules that apply to private parties, respectively—will

52. See descriptions, e.g., of Second Pair of Eyes Review pilots, described in the companion online appendix. Chien, *supra* note 17.

53. *See generally* Colleen Chien, *Piloting Applicant-Initiated 101 Deferral Through a Randomized Controlled Trial*, 2019 PATENTLY-O PATENT L.J. 1. A related proposal was advanced by Dennis Crouch & Robert P. Merges, *Operating Efficiently Post-Bilski by Ordering Patent Doctrine Decision-Making*, 25 BERKELEY TECH. L.J. 1673 (2010). To reduce the risks associated with the delays associated with deferral, the USPTO could put examination on an accelerated schedule. It could also enact a rule change with a sunset, as it did in the case of the Covered Business Method Program of the America Invents Act § 18, potentially supporting a "before and after" comparison.

54. *See* ZACH GUBLER, REGULATORY EXPERIMENTATION, FINAL REPORT 47–50 (2017).

55. *See* Kawaoka v. City of Arroyo Grande, 17 F.3d 1227, 1240 (9th Cir. 1994) (refusing to sustain an equal protection claim based on a temporary, but universally applied temporary water moratorium on the basis that "there is no evidence in the record that property similar to the [plaintiff's] was treated differently").

56. Because any difference in treatment can be understood as natural result of a free choice, rather than an arbitrary or mandatory governmental scheme, see Middleton v. Tex. Power & Light Co., 249 U.S. 152, 160 (1919) (holding that the consequences of not enrolling in an optional public insurance scheme did not violate equal protection because they were "the natural and inevitable result of a free choice, and not . . . legislative discrimination").

depend on the particulars. For the reasons described below, agency experiments that randomize internal processes rather than outward-facing rules may be subject to less stringent procedural requirements. Other aspects of a pilot's design will also have implications for its legality. For example, any number of factors—for example, that the difference in treatment does not lead to a justiciable claim or injury-in-fact,[57] that the pilot does not represent "final" agency action,[58] or that the pilot is mandated by Congress[59]—can reduce or eliminate certain legal risks under the Constitution or APA.

However, such design choices can also limit learning. For example, while including only voluntary participants in an experiment removes a number of concerns, it can also compromise the extent to which an experimental finding is generalizable, or its "external validity" if members of the test population differ in important ways from the regulated population, or behave differently because they know they are in an experiment.[60] As such it is important to understand the standards by which the most legally challenging types of rigorous pilots—pilots that randomize the rules or laws that govern the behavior of private parties—have been and are likely to be adjudicated in federal courts. Although Constitutional and APA claims will both depend in part on a pilot's "rational" or "reasoned" basis, leading at least one appellate court to conclude that "the equal protection argument can be folded into the APA argument,"[61] this Part separates them for the sake of completeness.

While few cases have directly considered rigorous pilots, the logic that courts have applied to randomized governmental behavior, and to experimentation and experimental evidence more generally provide windows into how courts are likely to decide challenges to such pilots. The paragraphs below consider cases involving all three types of pilots.

---

57. Injury that is "concrete and particularized" is required for Article III, Section 2 federal judicial standing. Spokeo, Inc. v. Robins, 136 S. Ct. 1540, 1545 (2016); *see also* Friends of Animals v. U.S. Fish & Wildlife Serv., 879 F.3d 1000, 1003 (9th Cir. 2018) (refusing to find standing in the case of animal advocacy organization that alleged harm based on the killing of one species of bird to conserve another as part of an experimental EPA program because the organization could not show that any of its members planned to visit the area where the birds were expected to experience adverse impacts). Even when injury-in-fact can be shown, causation and redressability may present additional hurdles.

58. As is required for review under the Administrative Procedure Act. *See* Bennett v. Spear, 520 U.S. 154, 178 (1997) (holding agency action to be "final" only when it (1) marks "the consummation of the agency's decision-making process," and (2) when "'rights or obligations have been determined' or 'from which legal consequences will flow'").

59. Potentially qualifying the agency action to being "committed to agency discretion by law" and outside the scope of judicial review. 5 U.S.C. § 701(a) (2012).

60. Called the "Hawthorne effect."

61. Nazareth Hosp. v. Sec'y U.S. Dept. of Health & Human Servs., 747 F.3d 172, 180 (3d Cir. 2014) (quoting Ursack Inc. v. Sierra Interagency Black Bear Grp., 639 F.3d 949, 955 (9th Cir. 2011) (internal quotations omitted)).

A. CONSTITUTIONAL CHALLENGES TO RIGOROUS POLICY PILOTS

The Equal Protection Clause forbids States from "deny[ing] to any person within its jurisdiction the equal protection of the laws,"[62] and enshrines the principle that like persons deserve to be treated alike.[63] The Fifth Amendment prohibits the Federal government from depriving persons "of life, liberty, or property, without due process of law"[64] and has been held to import an equal protection requirement equivalent to that established by the Fourteenth Amendment.[65] The concepts of equal protection and due process both "stem[] from [the] American ideal of fairness."[66] Rigorous pilots that impose work program requirements on some individuals eligible for welfare but not others, that put similarly situated debtors on different fee schedules, or that entitle some but not others for early release from prison at first glance might appear to fail the basic test. However, as described below, courts considering these and related experiments have reached the opposite conclusion, upholding their constitutionality.

1. The Interests Rationally Furthered by Policy Pilots

The Supreme Court has held that in areas of social or economic policy, a statutory classification that neither burdens fundamental rights nor targets a suspect class is consistent with Equal Protection as long as it "bears a rational relation to some [independent and] legitimate [legislative] end."[67] Rational basis review also applies when a party lodges a due process challenge to legislative action on the basis that it deprives persons of life, liberty, or property.[68] Generally speaking, pilots that do not "proceed along suspect lines"[69] such as race, alienage, or gender[70] should not be subject to heightened review merely based on the way they distinguish between participants and non-participants. In fact, rigorous pilots that randomly select certain individuals to receive a treatment arguably do not discriminate in the first place, but rather provide all with an *equal chance* of receiving the treatment. At least one lower court has endorsed this principle, stating in its review of a random ballot voting scheme, that "[t]here can be no denial of equal protection when *all share an equal opportunity*."[71]

---

62. U.S. CONST. amend. XIV.

63. City of Cleburne v. Cleburne Living Ctr., 473 U.S. 432, 439 (1985).

64. U.S. CONST. amend. V.

65. Weinberg v. Wiesenfeld, 420 U.S. 636, 638 (1975) ("[The] approach to Fifth Amendment equal protection claim [is] precisely the same as to equal protection claims under the Fourteenth Amendment.").

66. Bolling v. Sharpe, 347 U.S. 497, 499 (1954).

67. Romer v. Evans, 517 U.S. 620, 631 (1996).

68. Exxon Corp. v. Governor of Md., 437 U.S. 117, 125 (1978) (interpreting the Fifth Amendment).

69. FCC v. Beach Commc'ns, Inc., 508 U.S. 307, 313 (1993).

70. *See* City of Cleburne v. Cleburne Living Ctr., Inc., 473 U.S. 432, 440 (1985).

71. Campbell v. Bd. of Educ., 310 F. Supp. 94, 102–05 (E.D.N.Y. 1970) (emphasis added).

The types of fundamental rights that are protected by due process[72] are also not generally within the ken of executive agencies. Agency pilots that provide benefits to certain individuals do not "deprive" others of life, liberty, or property. Regarding an inmate's potential early release from prison through a test program, a lower court has concluded "the Petitioner does not have a constitutionally protected liberty interest in participating in the Pilot Program."[73] Most policy pilots will be reviewed under the permissive, rational basis standard.

In a post-*Lochner* world, courts have consistently held that the judiciary cannot "second guess the legislature on the [underlying] factual assumptions or policy considerations"[74] when applying rational basis review. Although legislatively-mandated policy pilots are arguably on stronger ground than agency-initiated pilots, courts have consistently applied the same, exceedingly forgiving rational basis test to a wide variety of forms of government action[75] including executive agency action.[76]

Why experiment? Courts generally recognize that pilots that treat similar parties differently rationally advance a variety of government interests. In the leading case on agency experimentation, *Aguayo v. Richardson*, the Second Circuit was asked to evaluate the constitutionality of a policy experiment approved by the Secretary of Health that imposed work requirements on a subset of individuals eligible for welfare benefits.[77] In upholding the Secretary's approval, the court held the "purpose to determine whether and how improvements can be made in the welfare system . . . [to be] as 'legitimate' or 'appropriate' as anything can be."[78]

In *In re Prines*,[79] petitioners brought an equal protection challenge to a pilot program that required them to pay debtor fees immediately in order to fund the program while those in non-pilot districts would not be charged until a year after their districts joined the pilot. Denying their claim, the Eighth

72. For a short and useful history and accounting of the rights that the Supreme Court has held to be fundamental, see Williams v. King, 420 F. Supp. 2d 1224, 1229–30 (N.D. Ala. 2006), *aff'd sub nom.* Williams v. Morgan, 478 F.3d 1316 (11th Cir. 2007).

73. Brown v. Rios, No. 08-5752, 2009 WL 5030768, at *11 (D. Minn. Dec. 14, 2009).

74. Sammon v. N.J. Bd. of Med. Exam'rs, 66 F.3d 639, 645 (3d Cir. 1995).

75. Koscielski v. City of Minneapolis, 435 F.3d 898, 902 (8th Cir. 2006) ("Due process claims involving local land use decisions must demonstrate the 'government action complained of is truly irrational . . . .'" (quoting Anderson v. Douglas County, 4 F.3d 574, 577 (8th Cir. 1993))); TriHealth, Inc. v. Bd. of Comm'rs, 430 F.3d 783, 788 (6th Cir. 2005) (finding the county board decision only unsustainable under equal protection if "irrational").

76. Cape May Greene, Inc. v. Warren, 698 F.2d 179, 184 (3d Cir. 1983) (citing lower court approval of EPA land use decision as "rational"); *see also* Schaeffler Grp. USA, Inc. v. United States, 786 F.3d 1354, 1369 (Fed. Cir. 2015) ("'[U]nder the due process and equal protection clauses,' agency action will be upheld 'if it has any conceivable rational basis.'" (alteration in original) (quoting California v. FCC, 905 F.2d 1217, 1238 (9th Cir. 1990))).

77. Aguayo v. Richardson, 473 F.2d 1090, 1109 (2d Cir. 1973).

78. *Id.*

79. U.S. Tr. v. Prines (*In re* Prines), 867 F.2d 478 (8th Cir. 1989).

Circuit recognized the legitimate government interest in establishing a nationwide self-supporting trustee system and found that "[t]he assessment of a . . . fee to fund the trustee program . . . satisfies the rational basis test."[80] Using similar reasoning, the Third Circuit found a pilot program that favored certain race-track operators over others Constitutionally legitimate because it furthered the legislature's objective of promoting the horse racing industry.[81] A lower court has upheld the legality of a test program on prisoners due to its purposes of vetting the feasibility of the program for reducing prison overcrowding.[82] To survive rational basis review, an agency's purpose merely needs to be a legitimate one, and the purpose of testing regulatory treatments seems to meet that standard.

These cases are consistent with decisions that have endorsed the rationality of "experimental" state action outside the strict confines of a pilot. For example, in *Moore v. Detroit Sch. Reform Bd.*, the Sixth Circuit considered a Michigan school reform law prompted by the perception of a "crisis" and "need for immediate action" on the part of the Michigan legislature. [83] Upholding the law, about which there was considerable uncertainty, the Sixth Circuit stated, "[s]tate legislatures need the freedom to experiment with different techniques to advance public education and this need to experiment alone satisfies the rational basis test."[84] The Pennsylvania Supreme Court applied similar reasoning to a narrowly targeted law aimed at the state's failing schools that, due to the large number of conditions placed on it, only applied to a single district, and put the law at risk of invalidation under a prohibition on "special legislation."[85] Upholding the law, the court held that the policy in question should be "assessed under the act's 'pilot program' before being made more generally available . . . there is nothing improper of this method of attacking social problems of statewide dimension, as the Legislature is free, for reason of necessity or otherwise, to address such issues incrementally."[86]

2. Unequal or Equal Treatment?

Do controlled experiments raise particular due process or equal protection concerns? In *Aguayo*, described above, the contested work program

80. *Id.* at 485 (citing United States v. Kras, 409 U.S. 434, 447–48 (1973)).

81. ACRA Turf Club, LLC v. Zanzuccki, 724 F. App'x 102, 111 (3d Cir. 2018).

82. Brown v. Rios, No. 08-5752, 2009 WL 5030768, at *10 (D. Minn. Dec. 14, 2009) ("The use of a test program on a limited number of [individuals] to measure the feasibility of achieving . . . legitimate [state] objectives is a reasonable means of proceeding." (citing O'Hara v. Rios, No. 08-5160, 2009 WL 3164724 (D. Minn. Sept. 28, 2009))).

83. Moore v. Detroit Sch. Reform Bd., 293 F.3d 352, 372 (6th Cir. 2002) (upholding the decision in Barefoot v. City of Wilmington, 37 F. App'x 626, 634 (4th Cir. 2002) to annex petitioner's land and stating that "the need to experiment to find the best procedure is itself a rational basis").

84. *Id.* (quoting Mixon v. Ohio, 193 F.3d 389, 403 (6th Cir. 1990)).

85. Harrisburg Sch. Dist. v. Zogby, 828 A.2d 1079, 1090 (Pa. 2003).

86. *Id.* at 1090–91 (internal citations omitted).

requirements applied to a fraction of the total eligible population.[87] In upholding the program, the Second Circuit held that the government interest in finding what worked to be "'suitably furthered' by a controlled experiment, a method long used in medical science which has its application in the social sciences as well."[88]

Courts have also generally rebuffed constitutional challenges to government actions based on their random distribution of burdens or benefits. In *Engquist v. Or. Dep't of Agric.*, the Supreme Court considered an Equal Protection challenge brought by a public employee who alleged that she was fired arbitrarily, while other similarly situated employees were not. In denying her "class-of-one" claim, the Court stated that "a random choice among rational alternatives does not violate the Equal Protection Clause."[89] A number of circuits have similarly held that a violation of equal protection requires discrimination that is intentional and not merely the product of "random government incompetence."[90]

These endorsements of *intentional* randomness complement court acceptance of randomness that arises out of *necessity*, for example when demand outstrips supply.[91] When a benefit or regulation is distributed to a limited group as part of a pilot, randomization in allocation prevents unfairness, giving all stakeholders an equal chance of receiving the treatment even if they don't all do so. However, while randomness is therefore tolerated, and even at times has been encouraged,[92] by the courts, it is not required.[93]

B. CHALLENGES TO EXPERIMENTAL RULEMAKING

Federal agencies must also worry about challenges to their actions under the APA. Whether or not an agency uses experimentation, its processes must be procedurally adequate and withstand judicial review.[94] As described below, courts will consider both the form (e.g., does the pilot involve a procedural

87. Aguayo v. Richardson, 473 F.2d 1090, 1109 (2d. Cir. 1973) (describing the application of work requirements to 25% and 2.5% of the 25%, or 0.625%, of the eligible population).

88. *Id.* at 1109.

89. Engquist v. Or. Dep't of Agric., 553 U.S. 591, 613 (2008) (Stevens, J., dissenting).

90. Wilson v. Northcutt, 441 F.3d 586, 591 (8th Cir. 2006) (quoting Batra v. Bd. of Regents of Univ. of Neb., 79 F.3d 717, 722 (8th Cir. 1996)).

91. Schenck v. City of Hudson, 114 F.3d 590, 593–95 (6th Cir. 1997) (holding the City's cap on development allotments to be consistent with substantive due process because it bore a rational relationship to the City's legitimate interest in maintaining sustainable growth and that a "lottery system is certainly a rational means of distribution because it avoids beauty contests . . . and is more efficient . . . to administer").

92. *See* Adam M. Samaha, *Randomization in Adjudication*, 51 WM. & MARY L. REV. 1, 44–45 (2009).

93. United States v. Claiborne, 870 F.2d 1463, 1467 (9th Cir. 1989) (finding that the Constitution does not require randomness).

94. For a summary of the requirements imposed by the APA on agency action, see generally TODD GARVEY, CONG. RESEARCH SERV., R41546, A BRIEF OVERVIEW OF RULEMAKING AND JUDICIAL REVIEW (2017), *available at* https://fas.org/sgp/crs/misc/R41546.pdf.

or interpretive rule) and the consequences of a policy pilot when reviewing the relevant administrative record. While conducting a policy pilot or forming a final rule based on an experiment can raise distinct issues, each experimental action of an agency, just like each non-experimental action of an agency, must be evaluated on its own merits.

1. Administering Policy Pilots

Agency statements "designed to implement, interpret, or prescribe law or policy or describing the organization, procedure, or practice requirements of an agency" are considered rules governed by the APA.[95] The APA's informal rulemaking procedures require agencies to provide the public with adequate notice of a proposed rule and enable interested persons a reasonable and meaningful opportunity to participate in the rulemaking process.[96] Announcements of adopted rules must be accompanied by "a concise general statement of their basis and purpose,"[97] and the agency's responses to "significant" comments.[98]

Whether or not a pilot is exempt from notice and comment requires a case by case determination,[99] and agencies have exercised discretion accordingly. However, pilots that "have considerable impact on ultimate agency decisions" or "substantially affect[] the rights of those over whom the agency exercises authority,"[100] even those that merely vary agency procedure,[101] require notice-and-comment. Adopting a rule on an experimental, then permanent basis can require the agency to go through an additional round or rounds of public comment.[102] But in light of the high failure rate of experiments,[103] it will often be the case that a practice is abandoned after a pilot has run.

While a voluntary, limited or temporary pilot will have less impact than a mandatory, agency-wide, or permanent practice, there's also a risk that the

---

95. 5 U.S.C. § 551(4) (2012). While an agency may use experimental techniques to carry out tactical or behavioral interventions, for example, in order to increase enrollment in savings plans, non-"legislative" practices are not subject to APA review.

96. *Id.* § 553(b).

97. *Id.* § 553(c).

98. Perez v. Mortg. Bankers Ass'n, 135 S. Ct. 1199, 1203, 1206–07 (2015).

99. *See* GUBLER, *supra* note 54, at 34 (discussing the applicability of 5 U.S.C. § 553(b)(3)(B)'s informal rulemaking procedures "good cause exception" to pilot processes). However, if the policy is a binding rule, it will need to be published in the Federal Register. 5 U.S.C. § 552(a).

100. Pickus v. U.S. Bd. of Parole, 507 F.2d 1107, 1113–14 (D.C. Cir. 1974).

101. Pilots that merely vary agency procedure and would otherwise be exempt from notice and comment under 5 U.S.C. § 553(b)(3) (holding that "interpretative rules, general statements of policy, or rules of agency organization, procedure, or practice," or situation that implicate "other good cause" are exempt).

102. GUBLER, *supra* note 54, at 10–11.

103. *See infra* Section III.B.

uncertainty associated with a temporary program will unsettle stakeholders and result in more speculative and premature reactions than would be elicited by a more formed proposal.[104] To reduce these risks, an agency's disclosure should at least inform the public of: the motivation for the pilot, the policy to be tested out, the information sought, and (where available) the agency's commitment to action based on the results of the pilot—a best practice in experimentation called "pre-commitment." Rulemaking based on experimental evidence may require more disclosure. The *Portland Cement*[105] decision of the D.C. Circuit and its progeny have been read to suggest "an obligation to reveal agency data from the outset,"[106] in order to support review and engagement with the data when dispositive.[107] In cases where the data is conclusive, even if more work to disclose, with the shared experience comes a greater chance of a common, ground truth, at least that is the hope.[108] As long as the agency faithfully follows the ordinary notice and comment process, procedural obstacles should be surmountable.

2. Judicial Review of Experimental Agency Action

The APA requires courts to set aside reviewable agency action that is "arbitrary, capricious, an abuse of discretion, or otherwise not in accordance with law"[109] or which otherwise exceeds the agency's constitutional or statutory authority.[110] Under *State Farm*, this requires the agency to "examine the relevant data and articulate a satisfactory explanation for [the] action including a 'rational connection between the facts found and the choice made.'"[111] Courts are not permitted to substitute the agency's judgment with their own,[112] but rather are required to ensure that the agency's explanations are satisfactory and do not run counter to the evidence.[113]

When an agency is implementing a policy pilot pursuant to a statutory authorization, judicial review will necessarily involve consideration of Congress' objectives and intents to ensure that an agency has not

104. GUBLER, *supra* note 54, at 36–38 (discussing the hypothetical possibility that rules introduced as experiments will elicit additional opposition, including by those who question its temporary nature).

105. Portland Cement Ass'n v. Ruckelshaus, 486 F.2d 375 (D.C. Cir. 1973), *cert. denied*, 417 U.S. 921 (1974).

106. PETER L. STRAUSS, ADMINISTRATIVE JUSTICE IN THE UNITED STATES 322 (3d ed. 2016).

107. In certain situations, an agency may be able to point to the open data they have made available generally to meet such an obligation.

108. A more cynical perspective might be that data will just create more issues to fight about.

109. 5 U.S.C. § 706(2)(A) (2012).

110. *Id.* §§ 701–706.

111. Motor Vehicle Mfrs. Ass'n v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 43 (1983) (quoting Burlington Truck Lines, Inc. v. United States, 371 U.S. 156, 168 (1962)).

112. *Id.* at 30.

113. *Id.* at 43.

> relied on factors which Congress has not intended it to consider, entirely failed to consider an important aspect of the problem, [or] offered an explanation for its decision that runs counter to the evidence before the agency, or is so implausible that it could not be ascribed to a difference in view or the product of agency expertise.[114]

While *State Farm* applies to agency action generally, not just agency piloting, the reversible nature of a pilot may bear upon the court's factual evaluation and cause a court to somewhat soften its hard look, at least, relative to a permanent irreversible version of the policy.

For example, in the *Aguayo* case described earlier, the Secretary's approval of a pilot program was made pursuant to 42 U.S.C. § 1315, which authorizes statutory waivers to support welfare demonstration projects. In reaching its decision, the Second Circuit considered the statute's objectives, the Secretary's consideration of the contested programs in view of the objectives and evidence, and the adequacy of the administrative record, including the agency's consideration and response to objections raised by the appellant.[115] But the court also relied upon "the fact that the programs were of limited duration and would remain under . . . on-going supervision"[116] to endorse the setting of "a lower threshold for persuasion" for the approval of experimental, temporary programs.[117] Finding there to be "no 'clear error of judgment,'"[118] the court specifically noted that "[e]xperience will be the best test of the reality of appellants' fears."[119]

While this single case by no means endorses a general heightened level of deference for policy pilots, it does underscore the fact-intensive nature of judicial review and the willingness of courts to engage in experimental agency action on its particular merits.

3. Judicial Deference or Indifference to Experimental Evidence?

What about agency rules that rely upon experimental evidence?[120] The D.C. Circuit has articulated some level of heightened deference to experimental evidence. In a 2002 decision involving a two-year experimental waiver on capacity ceilings by the Federal Energy Regulatory Commission, the Circuit began its analysis by citing "the special deference due agency experiments" and reiterating the Commission's statement, that, "[n]o matter how good the data suggesting that a regulatory change should be made, there

---

114. *Id.*

115. Aguayo v. Richardson, 473 F.2d 1090, 1106 (2d Cir. 1973).

116. *Id.*

117. *Id.* at 1103.

118. *Id.* at 1106 (quoting Citizens to Pres. Overton Park, Inc. v. Volpe, 401 U.S. 402, 416 (1971)).

119. *Id.*

120. The amount of deference owed to an agency's empirical conclusions was arguably unsettled by *Bus. Roundtable v. SEC*, 647 F.3d 1144 (D.C. Cir. 2011). For a discussion of the case, see GUBLER, *supra* note 54, at 30–33.

is no substitute for reviewing the actual results of a regulatory action."[121] The court went on to describe its practice of, "[f]or at least 30 years . . . [giving] special deference to agency development of such experiments, precisely because of the advantages of data developed in the real world."[122]

However, the arguably more defensible way of regarding experimental data is to treat it like any other form of evidence—on its own merits, in its own individual context. Rigorous experimental data *can* and often *will be* superior to other forms of data and deserve more deference. This appeared to be the case in *Christ the King Manor, Inc. v. Sec'y U.S. Dep't of Health & Human Servs.*, in which the Third Circuit had occasion to evaluate the Secretary of Health and Human Services' reliance on "actual-experience data" that showed the impact of a proposed rate change on the quality of Medicaid care.[123] Because the experiential data was "clearly more recent and more accurate than the [older] predictive data," the court found, it made little sense for the Secretary to rely on the older data and "ignore the 'answers to the test.'"[124] But there will be situations in which experimental data will *not* deserve special deference or elevation above other sources of data. As discussed in the next Section, any number of defects including methodological issues, unobservable factors and outcomes, changes in conditions, or conflicting or inconclusive results can limit what can be learned from a pilot, however experimental.

An indifferent, rather than deferential, posture is supported by a number of decisions that analytically separate *State Farm* into two distinct inquiries, one far more important than the other.[125] The first is to ascertain the data considered by the agency for its relevance, and the second, critical step, is to consider whether or not the agency's choice or action rationally follows from them. The way that courts have regarded non-experimental evidence associated with the first step—as imperfect, fallible, and in most cases, adequate—suggests they would apply a similar approach to experimental evidence. For example, several appellate courts have held that the consideration of older[126] or second-best[127] data by itself, is not arbitrary and

121. Interstate Nat. Gas Ass'n of Am. v. FERC, 285 F.3d 18, 30 (D.C. Cir. 2002).

122. *Id.*

123. Christ the King Manor, Inc. v. Sec'y U.S. Dep't of Health & Human Servs., 673 F. App'x 164, 167, 172 (3d Cir. 2016).

124. *Id.* at 172.

125. Namely, to "examine the relevant data and articulate a satisfactory explanation for [the] action including a 'rational connection between the facts found and the choice made.'" Motor Vehicle Mfrs. Ass'n v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 43 (1983) (quoting Burlington Truck Lines, Inc. v. United States, 371 U.S. 156, 168 (1962)).

126. *See* Fishermen's Finest, Inc. v. Locke, 593 F.3d 886, 898 (9th Cir. 2010) (finding the use of data from 1995–2003, rather than 2004–2005 which was impracticable to include, neither arbitrary nor capricious). *See generally* Miss. Comm'n on Envtl. Quality v. EPA, 790 F.3d 138 (D.C. Cir. 2015) (finding the EPA's use of older data not arbitrary or capricious).

127. Dist. Hosp. Partners, L.P. v. Burwell, 786 F.3d 46, 56 (D.C. Cir. 2015).

capricious.[128] In a similar vein, the D.C. Circuit has ruled that anecdotal observations can suffice, as long as the inferences drawn from them are reasonable. In the case of *Nat'l Black Media Coal. v. FCC*, for example, it blessed "the Commission's decision to forego a rigorous statistical evaluation," in view of the modest inferences it made from its analysis, to be reasonable.[129] There is no affirmative obligation to experiment.[130]

But what courts have been unwilling to tolerate are misuses of data to reach unsupported conclusions. Selectively relying on portions of a study while ignoring other, contrary parts of the study has been held to be unreasonable.[131] So has willfully ignoring new and better data within an agency's grasp,[132] though whether data are "better" is often closer to contestable opinion than objective fact. In addition, while an agency's embrace of experimentation can demonstrate that the agency is considering the effect of its policy, it does not necessarily follow that the subsequent policy developed by the agency is reasonable, the Ninth Circuit has found.[133]

C. CONCLUSION

Collectively, these cases confirm that the Constitution and APA present limited if any special barriers to rigorous piloting. No decision that I could find has held that experimentation rendered an agency rule per se arbitrary and capricious because it demonstrated a lack of sufficient information to select a course of action. To the contrary, courts are generally accepting of experimentation as a rational way of learning about a policy and value

128. *But see generally* Ctr. for Biological Diversity v. Zinke, 900 F.3d 1053 (9th Cir. 2018) (holding FWS's failure to rely on best scientific and commercial data available was arbitrary and capricious).

129. Nat'l Black Media Coal. v. FCC, 706 F.2d 1224, 1228 (D.C. Cir. 1983) (describing how the Commission noted that with more resources it would have used a "more scientific method" but that such rigor was not required, where the only conclusions drawn from the data were that exempt television applications were denied less than half than non-exempt stations and that "exempt stations were doing at least as satisfactory a job of providing responsive programming as were non-exempt stations," "determinations that were 'primarily of a judgmental or predictive nature'" (quoting FCC v. Nat'l Citizens Comm. for Broad., 436 U.S. 775, 813 (1978))).

130. Am. Whitewater v. Tidwell, 770 F.3d 1108, 1116 (4th Cir. 2014) ("Where the agency's conclusion otherwise rests on a firm factual basis, nothing in the APA requires it to experiment with a practice . . . .").

131. *See generally* Genuine Parts Co. v. EPA, 890 F.3d 304 (D.C. Cir. 2018) (holding that the EPA acted arbitrarily and capriciously by relying upon portions of studies that supported its position, while ignoring cross sections in those studies that did not).

132. Am. Tunaboat Ass'n v. Baldrige, 738 F.2d 1013, 1016 (9th Cir. 1984) (finding the "[National Oceanic & Atmospheric Administration's] decision [to] ignore[] a comprehensive database that [was] the product of many years' effort by trained research personnel" to be arbitrary and capricious).

133. Cal. Energy Comm'n v. Bonneville Power Admin., 909 F.2d 1298, 1310 (9th Cir. 1990). The Ninth Circuit cited the BPA's decision to experiment as evidence that it was giving impacts of its policy significant consideration. *Id.* "This conclusion does not end the inquiry, however. We must now address whether BPA's policy is reasonable." *Id.*

randomization as an adequate, and potentially superior,[134] way of distributing a benefit or a burden to a population.

At the same time, as described above, rigorous experimentation will neither remove the burdens of ensuring an adequate nexus to a legitimate governmental objective, nor absolve an agency of procedural compliance with notice and comment, nor shield agency from scrutiny. When the stakes are high and agency action is likely to have material consequences for stakeholders, experimental evidence, like other evidence, is likely to be contested and challenged. While on one hand, experimental evidence can create a shared experience, on the other hand, stakeholders and even agencies are likely to interpret evidence in the manner that advances their interests.[135] Matters of institutional competence may come into play if the court is unsure how to evaluate an an agency's experimental design or interpretation of experimental results.

Still, for satisfying an agency's own learning agenda, and injecting greater certainty into its decisions, "hard" experimental evidence will often compare favorably. While feedback through litigation requires the courts to address petitioner's question, experimentation allows the policy- or law-maker to ask her question, as well as specify the method that will be used to answer it. Expert opinion evidence provided through lobbying and participation in notice-and-comment will in some cases be more partial and less representative than experimental evidence. But even if legal, in what situations is experimentation going to be practically feasible and desirable? The answers that agencies have given to this question forms the basis for the next Part.

III. (A FRAMEWORK FOR PROPOSING) RIGOROUS POLICY PILOTS (THAT) ARE FEASIBLE AND WORTHWHILE

The previous Part addresses legal challenges to rigorous policy pilots. However, significant practical and institutional challenges to rigorous piloting will often attend. Recent research has found that people tend to view as inappropriate pilots that purport to randomize a treatment even when the underlying treatment is unobjectionable,[136] suggesting that buy-in of the

134. For a proposal that endorses randomization as an antidote to historical bias in law enforcement, see Bernard E. Harcourt, Against *Prediction: Sentencing, Policing, and Punishing in an Actuarial Age* 15 (U. of Chi. Law Sch. Pub. Law & Legal Theory Working Papers, Working Paper No. 94, 2005), https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=1021&context=public_law_and_legal_theory (arguing that "[w]e would be better off as a society if we deployed our criminal justice measures more randomly").

135. For a scathing critique of the EPA's reliance on industry-provided simulations involving the herbicide dicamba that did not adequately capture what was actually happening in the field, despite warnings from scientists, see Liza Gross, *Scientists Warned This Weed Killer Would Destroy Crops. EPA Approved It Anyway*, REVEAL (Nov. 13, 2018), https://www.revealnews.org/article/scientists-warned-this-weed-killer-would-destroy-crops-epa-approved-it-anyway.

136. *See generally* Michelle N. Meyer et al., *Objecting to Experiments That Compare Two Unobjectionable Policies or Treatments*, 116 PNAS 10723 (2019).

relevant population, if required, will require education and communication. More upfront work is also required to make explicit the assumptions behind an intervention, and then to develop a strategy for testing them. Agencies carrying out rigorous tests must marshal resources, data, statistical expertise, support for departing from the status quo, and explanations for skeptics.[137] There often will be no agreed upon measure of success. Even when there is, attrition, changed circumstances, poor design or a host of other factors can lead to a lack of inconclusive or null results. The results may be rejected. For example, in their review of the 50 or so randomized experiments in U.S. law that have ever been conducted, mostly in the field of criminal law and judicial procedure, Greiner and Matthews found a major problem to be the refusal of judges and lawyers to accept experimental results that did not jive with their sense of the "right answer."[138]

However, as daunting as the hurdles to rigorous piloting may appear, they are not new. Although individually perhaps little known beyond each relevant agency or initiative and attendant stakeholder community, collectively, the experience of executive agencies with rigorous pilots is not insubstantial. Outside of medical contexts,[139] randomized controlled trials ("RCTs") were carried out in waves during the 1930s–1970s to address social policy questions, and then again in the 1990's to support the development and implementation of welfare and employment laws.[140] Between 2009 and 2011, Congress enacted six "tiered-evidence" programs at multiple agencies to fund replication RCTs to determine whether prior findings could be reproduced on a larger scale.[141] The Institute of Educational Sciences within the Department of Education,[142] Center for Medicare and Medicaid

137. *See, e.g.*, Eileen M. Ahlin, *Conducting Randomized Controlled Trials with Offenders in an Administrative Setting*, 36 AM. J. EVALUATION 164, 168–72 (2015) (discussing, in the context of a series of RCTs carried out by a state agency, the resistance of Administrators and staff over changes in leadership).

138. *See* Greiner & Matthews, *supra* note 35, at 303, 307 (internal quotation marks omitted) ("[O]ur speculation is that because lawyers train themselves in instrumental, as opposed to inquisitive, analysis, the result is ingrained arrogance. There is less willingness to acknowledge the kind of uncertainty needed to make an [sic] [randomized control trial] relevant.").

139. Such trials date back at least to the time of the Salk polio vaccine trials in the 1940s, as described in Marcia Meldrum, "*A Calculated Risk": The Salk Polio Vaccine Field Trials of 1954*, 317 BMJ 1233, 1233–36 (1998).

140. *See* Jon Baron, *A Brief History of Evidence-Based Policy*, 678 ANNALS AM. ACAD. POL. & SOC. SCI. 40, 40–50 (2018). Lawmakers have also provided waivers to states to implement federal policy and mandated evaluations in exchange. *See, e.g.*, Jessica Bulman-Pozen & Heather K. Gerken, *Uncooperative Federalism,* 118 YALE L.J. 1256, 1274 (2009) (describing the authorization of statutory waivers for pilot or demonstration projects).

141. *See generally* RON HASKINS & GREG MARGOLIS, SHOW ME THE EVIDENCE: OBAMA'S FIGHT FOR RIGOR *AND* RESULTS IN SOCIAL POLICY (2015) (describing trials at HHS, the Department of Education, Corporation for National and Community Service, and Department of Labor).

142. *About IES: Connecting Research, Policy and Practice*, IES, https://ies.ed.gov/aboutus (last visited June 8, 2019) (describing as one of its six missions, to support rigorous testing of new approaches "through pilot studies and rigorous testing at scale"). For an account of the IES as

Innovation within the Centers for Medicare & Medicaid Services,[143] and Office of the Chief Evaluation Officer ("CEO") within the Department of Labor ("DOL"),[144] each support rigorous evaluation.[145] Various Presidents have also called for rigorous controlled testing of policies.[146]

It is also worth remembering that calls for rigorous evaluation are not new and their results have not necessarily been encouraging. Critics have described evaluation mandates embedded in the Government Performance and Results Act of 1993 ("GPRA"), the GPRA Modernization Act of 2010 ("GPRAMA"), and successive generations of retrospective review requirements as hampered by resource constraints, pro forma compliance, and a lack of oversight.[147]

---

part of the broader move away from the "overly politicized" and fad-driven development of US education, see Benjamin Michael Superfine, *New Directions in School Funding and Governance: Moving from Politics to Evidence*, 98 KY. L.J. 653, 686 (2009).

143. The Innovation Center was created for the purpose of "testing 'innovative payment and service delivery models to reduce program expenditures . . . while preserving or enhancing the quality of care . . . [including through] rigorous evaluation of the impact of each model on outcomes of interest." *About the CMS Innovation Center,* CMS.GOV, https://innovation.cms.gov/About (last updated May 14, 2019) (alteration in original) (quoting 42 U.S.C. § 1315a (2012)).

144. *See generally* U.S. DEP'T OF LABOR, CHIEF EVALUATION OFFICE, FY2018 PLAN FOR USE OF SET ASIDE (2018), *available at* https://www.dol.gov/asp/evaluation/completed-studies/CEO-FY-2018-Evaluation-Plan.pdf (describing Youth Connect and Ready to Work random trials, among a large number of evaluations, most multi-year).

145. Under the 2018 Evidence Act, each of 24 major agencies is required to follow the DOL model and have its own CEO. Foundations for Evidence-Based Policy Making Act of 2018, Pub. L. No. 115-435, § 313, H.R. 4174 (2019); *see also* OPRE, https://www.acf.hhs.gov/opre (last visited June 10, 2019) (describing the office's work as carrying out "rigorous research and evaluation projects . . . . includ[ing] program evaluations, research syntheses and descriptive and exploratory studies").

146. *See, e.g.*, Executive Order 13,707 issued by President Obama, encourages agencies to "rigorously test and evaluate the impact" of policies that leverage behavioral science insights. Exec. Order No. 13,707, 3 C.F.R. § 13707 (2016). Executive Order 13,801, issued by President Trump, gives a preference to "multi-site randomized controlled trials," in the evaluation of job training programs Exec. Order No. 13,801, 82 Fed. Reg. 28,229 (June 15, 2017). Predating them both, the Office of Management and Budget blessed the use of randomized control trials as the "most effective method of evaluation." *What Constitutes Strong Evidence of a Program's Effectiveness*, OFFICE OF MGMT. & BUDGET 2004, *available at* https://obamawhitehouse.archives.gov/sites/default/files/omb/part/2004_program_eval.pdf (identifying well-conducted RCTs as the most effective method of evaluation).

147. *See* Seth Harris, *Managing for Social Change: Improving Labor Department Performance in a Partisan Era*, 117 W. VA. L. REV. 987, 1004–11 (2015) (describing how, in implementing the Government Performance and Results Act of 1993 ("GPRA") and the GPRA Modernization Act of 2010 ("GPRAMA"), executive branch entities "were able to get away with poor performance and pro forma . . . compliance because they knew Congress was not paying attention"); *see also* GOV'T ACCOUNTABILITY OFFICE, MEDICAID DEMONSTRATIONS: EVALUATIONS YIELDED LIMITED RESULTS UNDERSCORING NEED FOR CHANGES TO FEDERAL POLICIES AND PROCEDURES (Jan. 2018), *available at* https://www.gao.gov/products/GAO-18-220 (describing the shoddiness of Section 1115 Medicare evaluations due to the lack of enforcement of the requirements of rigor). For an overview of the criticisms of regulatory lookback efforts, see ACUS Adoption of Recommendations, 79 Fed. Reg. 75,114, 75,115–16 (Dec. 17, 2014) (describing criticisms of

However, several developments make it an opportune time to persevere and attempt to answer, not abandon their call. First the costs of controlled experimentation are going down, thanks to the advance of open data,[148] and the adoption of "agile" and "policy lab" approaches in government,[149] which have increased the expertise and decreased measurement and assessment costs. Second, the embrace of performance- or results-based lawmaking has shifted attention away from policy processes and towards policy outcomes, discoverable through trial and error or disciplined experimentation.[150] Third, the 2018 Evidence Act, if implemented as intended, should significantly shore up agency capacity for rigorous evaluation. The Act requires departments and agencies within the Act's jurisdiction to submit annual plans for identifying and addressing policy relevant questions that specify "data to . . . facilitate the use of evidence in policy making."[151] It establishes an Interagency Council on Evaluation Policy, requires each relevant agency to designate a Chief Evaluation Officer and requires the Office of Management and Budget ("OMB") to establish an Advisory Committee on Data for Evidence Building to advise on expanding access to and use of federal data for evidence building.[152] These activities should inform and ease at least two of the steps of rigorous piloting described below—identifying important questions and developing the evidence to evaluate them.

---

regulatory lookback efforts including conformance with the Regulatory Flexibility Act and various generations of retrospective review as resource-constrained, "inherently deregulatory," and potentially pro forma).

148. Catalyzed during the Obama Administration by "M-13-13." *See* Memorandum from Sylvia Burwell et al. (May 9, 2013), *available at* https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf.

149. For example, as advanced by the US Digital Service of the Office of Management and Budget, created in 2014. *See generally Quick and Dirty Skinny on Agile Software Development*, U.S. DIG. SERV., *available at* https://techfarhub.cio.gov/assets/files/Quick%20and%20Dirty%20Skinny%20on%20Agile%20Software%20Development-I2.pdf (providing information on agile software development strategies). For a map of some government labs, see PARSONS DESIS LAB, GOV INNOVATION LABS CONSTELLATION 1.0, at 1–2 (2013), http://nyc.pubcollab.org/files/Gov_Innovation_Labs-Constellation_1.0.pdf. Government labs can operate at the local (for example, New York City's iZone), state (for example, the Rhode Island Innovative Policy Lab), and national (for example, UK Justice Data Lab) levels. *See id.*; Ministry of Justice, *Official Statistics: Justice Data Lab Statistics: January 2018*, GOV.UK (Jan. 11, 2018), https://www.gov.uk/government/statistics/justice-data-lab-statistics-january-2018.

150. Lauren E. Willis, *Performance-Based Consumer Law*, 82 U. CHI. L. REV. 1309, 1311 (2015) (describing the use of performance-based regulation monitored through field testing in environmental regulation, wherein "rather than the law dictating that a factory smokestack must incorporate a particular scrubber, the law sets limits on a firm's emissions and the firm can then determine how to meet those limits").

151. Evidence Act of 2018, Pub. L. No. 115-435, § 312(a), H.R. 4174 (2019).

152. *Id.* § 315.

Finally, many agencies are already actively engaging in piloting.[153] While rigorous designs are by no means part of the standard repertoire,[154] designing policies with identification strategies in mind can improve the chances that the actual impact of law or policy on outcomes will be discernible.

Can practical obstacles to rigorous piloting be overcome and when is it worthwhile to do so? The next Sections draw on the past experiences of agencies, including the USPTO, to illustrate that rigorous pilots are not only feasible but can be informative and useful. They present a framework for proposing pilots that comprises specifying (1) questions that matter; (2) authority and agency resources; (3) theory of change; (4) testing protocol; (5) evidence and (6) resources ("MATTER").

A. STEP 1 ("M"): SELECT A QUESTION THAT MATTERS

The types of questions that rigorous policy pilots have addressed in the past provide some insight into the questions that could be feasible and worthwhile to be addressed in the future. They span diagnostic questions (how accurate is the Trademark registry?,[155] asked by the USPTO), policy development questions (what is the impact of price-test restrictions on short

153. *See, e.g.*, Ensuring Program Uniformity at the Hearing and Appeals Council Levels of the Administrative Review Process, 81 Fed. Reg. 90,987, 90,992 (Dec. 16, 2016) (to be codified at 28 C.F.R. pts. 404, 405, 416); Small Business Government Contracting and National Defense Authorization Act of 2013 Amendments, 81 Fed. Reg. 34,243, 34,252 (May 31, 2016) (to be codified at 13 C.F.R. pts. 121, 124–127) (evaluating the Small Business Teaming pilot); Accreditation of Third-Party Certification Bodies to Conduct Food Safety Audits and to Issue Certifications, 80 Fed. Reg. 74,570, 74,583 (Nov. 27, 2015) (21 C.F.R. pts. 1, 11, 16) (discussing the USDA Agricultural Marketing Service pilot); Foreign Supplier Verification Programs for Importers of Food for Humans and Animals, 80 Fed. Reg. 74,226, 74,327 (Nov. 27, 2015) (to be codified at 21 C.F.R. pts. 1, 11, 111) ("We are transitioning the systems recognition program from the pilot phase to the implementation phase."); Version 5 Critical Infrastructure Protection Reliability Standards, 78 Fed. Reg. 72,756, 72,760 (Dec. 3, 2013) (to be codified at 18 C.F.R. pt. 40) (introducing the National Energy Regulatory Council pilot program); Attestation Process for Employers Using F–1 Students in Off-Campus Work, 78 Fed. Reg. 69,538, 69,538 (Nov. 20, 2013) (to be codified at 20 C.F.R. pt. 655) (referring to the off-campus work F-1 student pilot program); Ability-to-Repay and Qualified Mortgage Standards Under the Truth in Lending Act (Regulation Z), 78 Fed. Reg. 35,430, 35,436 (June 12, 2013) (to be codified at 12 C.F.R. pt. 1026) (discussing the Single Family Housing Guaranteed Rural Refinance pilot); Changes to Implement Micro Entity Status for Paying Patent Fees, 77 Fed. Reg. 75,019, 75,031 (Dec. 19, 2012) (to be codified at 37 C.F.R. pt. 1) (describing the LegalCORPS Inventor Assistance Pilot Program); *see also* ADMIN. CONFERENCE OF THE U.S., SSA DISABILITY BENEFITS ADJUDICATION PROCESS: ASSESSING THE IMPACT OF THE REGION I PILOT PROGRAM 1 (2013), https://www.acus.gov/sites/default/files/documents/Assessing%20Impact%20of%20Region%20I%20Pilot%20Program%20Report_12_23_13_final.pdf (discussing the Region 1 pilot a "laboratory of sorts to test and compare . . . theoretical propositions").

154. *Accord* Michael Abramowicz et al., *Randomizing Law*, 159 U. PA. L. REV. 929, 933 (2011) (noting the dearth of experiments in securities law or taxation).

155. Changes in Requirements for Affidavits or Declarations of Use, Continued Use, or Excusable Nonuse in Trademark Cases, 82 Fed. Reg. 6,259, 6,260 (Jan. 19, 2017) (to be codified at 37 C.F.R. pts. 2, 7) (providing overview of the 2012 announcement of a two-year randomized pilot program).

sales?,[156] asked by the SEC), long-term outcomes questions (do housing and services interventions for homeless families prevent them from returning to homelessness?,[157] asked by HUD), and policy implementation questions (what strategies work best for driving benefits uptake?, asked by multiple agencies[158]). It is hard to imagine other ways of getting evidence of comparable quality about the performance (as opposed to, say, the popularity) of legal and policy interventions.

But as wide-ranging as the pool of potentially good questions might be, the range of questions poorly suited for rigorous experimentation is also considerable. They include situations where priors, dictated for example by politics or ideology, are so strong that even high-quality evidence is unlikely to be persuasive, as well as situations where the importance of implementation and contextual factors make it difficult to replicate observed effects.[159] Situations where causal proof would be frivolous,[160] for which the important outcomes are not subject to measurement,[161] or where the observations would be too few perhaps because the questions are so big, encompassing imagining and measuring the impact of major changes in institutions, for example, also present poor candidates.

These criteria, in turn, make rigorous experimentation easier to carry out with respect to some agency actions than others. Agency interventions that are customized and non-repeatable, for example involving long-term investigations, are not easily tested through field RCTs. However, lab tests that vet mechanisms, address basic research questions, or simulate market dynamics can still be useful. The CFPB, which is tasked with taking enforcement actions, has used rigorous testing to increase its understanding of market mechanisms, for example, by lab-testing disclosures and simulating consumer environments and markets to explore how the complexity of a product's price influences the consumption decision.[162]

Do these constraints shrink the pool of questions addressable by rigorous pilots to the uncontroversial, transactional, and therefore marginal? Over

156. OFFICE OF ECON. ANALYSIS & U.S. SEC. & EXCH. COMM'N, ECONOMIC ANALYSIS OF THE SHORT SALE PRICE RESTRICTIONS UNDER THE REGULATION SHO PILOT 4 (2007), *available at* https://www.sec.gov/news/studies/2007/regshopiloto20607.pdf.

157. Office of Policy Dev. & Research, *The Family Options Study*, HUD USER, https://www.huduser.gov/portal/family_options_study.html (last visited June 10, 2019).

158. *Methods*, OFFICE OF EVALUATION SCI., *available at* https://oes.gsa.gov/methods.

159. For example, in the prevention of opioid deaths, as discussed in Christopher J. Ruhm, *Shackling the Identification Police?* 5–6 (Nat'l Bureau of Econ. Research, Working Paper No. 25320, 2018), *available at* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3294925.

160. For example, to demonstrate the effectiveness of parachutes as satirized in 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 (2003).

161. Underhill, *supra* note 28, at 11–12 (discussing measurement problems raised by fundamental values and other "unquantifiables").

162. Johnson & Leary, *supra* note 48, at 190.

several decades,[163] there has been a debate about the shortcomings of insisting on causal methods and their tendency to favor, "good answers instead of good questions."[164] As Raj Chetty has commented, "[p]eople think about the question less than the method . . . so you get weird papers, like sanitation facilities in Native American reservations."[165] But keeping at the center of the inquiry the informational deficits that only pilots can address, and how pilots fit into a broader learning and law-making agenda can ensure that experimentation is worth the candle.

For example, when the HHS started advancing its controlled welfare experiments for evaluating welfare policy, it focused on "the most significant questions about policy and practice,"[166] to get buy-in from relevant administrators and agency employees responsible for implementing the policy. The RCTs "did not assess reforms dreamed up by policy wonks . . . [but rather] bubbled up from governors, their staffs, and community activists —people with finely calibrated judgment on political timing."[167] Even though the resulting programs took years to design and implement, the work remained relevant to a wide variety of stakeholders. During the 1980's and 1990's, HHS funded or facilitated more than 85 welfare RCTs, "building a sizeable body of credible evidence with clear policy relevance."[168] Among the findings: Short-term job-search assistance and training had larger effects than remedial education and combining mandatory participation in employment-focused services with earning supplements was also effective at raising overall income and moving individuals out of poverty.[169] The results informed broader welfare lawmaking and policy development in the United States. Unemployment insurance ("UI") policy is another context in which rigorous pilots have been used to answer open question in law- and policy-making.[170]

As such, among policy actors, agencies enjoy certain comparative advantages (though not free reign), to generate information. Unlike courts, which are constrained by the parties and facts before them, pilot designers

163. *See* Ruhm, *supra* note 159, at 2–3 (reviewing a literature that dates back to 1995).

164. Joshua D. Angrist & Jörn-Steffen Pischke, *The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con Out of Econometrics*, 24 J. ECON. PERSP. 3, 6 (2010).

165. *Id.* at 24.

166. J.M. Gueron, *The Politics and Practice of Social Experiments: Seeds of a Revolution*, *in* HANDBOOK OF FIELD EXPERIMENTS 3 (Abhijit Vinayak Banerjee & Esther Duflo eds., 2017).

167. *Id.* at 57.

168. *Id.* at 45.

169. *Id.*

170. COUNCIL OF ECON. ADVISORS, ECONOMIC REPORT OF THE PRESIDENT—CHAPTER 7: EVALUATION AS A TOOL FOR IMPROVING FEDERAL PROGRAMS 278 (2014), *available at* https://www.govinfo.gov/app/details/ERP-2014/ERP-2014-chapter7/context (describing several rounds of testing, using randomized design, to probe whether UI reduced incentives to find work: A first set considered interventions that included in-person reviews focused on whether reemployment was effective, a second trial looked at a broader set of outcomes, and the impact of personalized reemployment services on them; the results informed development of the Middle Class Tax Relief and Job Creation Act of 2012).

are free, within the limits of their authority, existing law, and related considerations, to design and adapt suggested policy interventions. Picking questions that are relevant to all branches of government, but which exploit an agency's comparative advantages can be particularly generative. Pilots can also be used to vet competing claims from stakeholders. For example, in 2018, the USPTO proposed a new procedure for amending patents challenged in post-grant proceedings. One aspect of the new procedures that was highly contested among the 49 responses it received concerned the appropriate timeline for amendment. Rather than going back to the drawing board, the USPTO noticed a pilot version of the program, and noted that it reserved the right to amend its terms in response to feedback and experience generated during the pilot.[171]

B. STEP 2 ("A"): CONSIDER EXISTING AUTHORITY AND AGENCY RESOURCES

As discussed in Part II, the legal requirements for carrying out pilots related to an agency's objectives need not be particularly strenuous. But marshalling the necessary agency approvals can present major hurdles. Implementing pilots can require employees to vary their behavior or depart from the normal course of operations, introducing both time and disruption costs.

Pre-decisional consultation processes can allow labor and agency to reach agreement about the terms and conditions of a pilot.[172] For example, when the USPTO decided to pursue a pilot that would require unionized examiners to add more detail to the patent record, agency leadership and labor negotiated and conferred. The resulting memorandum of understanding that detailed eligibility criteria, how examiners would be invited to participate in the process, and how data would be shared by both labor and management to evaluate the pilot's success.[173]

Another way to justify rigorous pilots is to integrate them into existing compliance obligations. The Administrative Conference of the United States has recommended using rigorous evaluations to fulfill various retrospective review mandates using a framework that closely resembles the one discussed in this Essay.[174] When rigorous retrospective evaluations take place, the results

171. Notice Regarding a New Pilot Program Concerning Motion to Amend Practice and Procedures in Trial Proceedings Under the America Invents Act Before the Patent Trial and Appeal Board, 84 Fed. Reg. 9,497, 9,502, 9,504 (March 15, 2019).

172. *See* Exec. Order No. 13,522, 3 C.F.R. § 13522 (2010). Though the Executive Order ("EO") was subsequently rescinded by President Trump in late 2017, the issue of consulting with labor before implementing an experiment will remain present in some experiments. Exec. Order No. 13,812, 3 C.F.R. § 13812 (2018).

173. Memorandum of Clarity of the Record Pilot MOU between the Patent Office Professional Association and the U.S. Patent & Trademark Office (Feb. 4, 2016), *available at* https://www.uspto.gov/sites/default/files/documents/clarity-of-the-record-pilot-mou-signed.pdf.

174. ACUS Adoption of Recommendation, *supra* note 147, at 751,162 (outlining a framework for rigorous retrospective review).

are often surprising: The Drug Resistance Education programs ("D.A.R.E") of the 1990s and 2000s, implemented in more than 75% of U.S. schools,[175] wrap-around services for prisoners reentering society,[176] deworming,[177] and 90% of the interventions evaluated by the Institute for Education Sciences from 2002 to 2013[178] are among the interventions that have been found to be ineffective.

C. STEP 3 ("T"): IDENTIFY A TREATMENT AND THEORY OF CHANGE

While pilots can be used to evaluate already-implemented policies, prospectively proposing a policy pilot involves identifying an intervention for evaluation grounded in a theory of change. Evidence is likely to be most valuable and valued when it addresses deficits left by traditional information gathering processes. Tests can address competing stakeholder narratives and theories of change for example, or can be used to try out new ideas.[179] Within the USPTO, participants in the Office's "Edison Scholar" program have helped to develop pilot programs including the "Glossary pilot,"[180] other pilots have come from or been supported by the Public Patent Advisory Committee ("PPAC"),[181] a Federal Advisory Committee of the agency.

Regardless of source, it is critical to connect the identified treatment to the eventual intended outcome. By answering the question, "how is treatment X supposed to achieve outcome Y," a law- or policy-maker describes a "theory of change." A widely-accepted method of specifically testing the underlying theory of change is through "mechanism experiments," a technique for

175. H.R. REP. NO. 115-435, at 4–5 (2017).

176. Jennifer L. Doleac, *Wrap-Around Services Don't Improve Prisoner Reentry Outcomes*, 38 J. POL'Y ANALYSIS & MGMT. 508 (2019).

177. DAVID C. TAYLOR-ROBINSON ET AL., DEWORMING DRUGS FOR SOIL-TRANSMITTED INTESTINAL WORMS IN CHILDREN: EFFECTS ON NUTRITIONAL INDICATORS, HAEMOGLOBIN AND SCHOOL PERFORMANCE 2 (2012) (concluding, on the basis of a review of 42 trials, "[o]ur interpretation of this data is that it is probably misleading to justify contemporary deworming programmes based on evidence of consistent benefit on nutrition, haemoglobin, school attendance or school performance as there is simply insufficient reliable information to know whether this is so").

178. *Hearing on What Works / Evaluation Before the Subcomm. on Human Resources of the H. Comm. On Ways & Means*, 113th Cong. 3 (2013) (statement of Jon Baron, President, Coalition for Evidence-Based Policy), *available at* http://coalition4evidence.org/wp-content/uploads/2013/07/Testimony-before-Ways-and-Means-HR-subcommittee-7.17.13-Jon-Baron.pdf.

179. For a list of sources, see Thomas Kalil, *Policy Entrepreneurship at the White House: Getting Things Done in Large Organizations*, 11 INNOVATIONS 4, 18 (2017).

180. Described at *Glossary Initiative*, USPTO, https://www.uspto.gov/patent/initiatives/glossary-initiative (last visited June 17, 2019) and seeded by the work of Professor Peter Menell during his time at the USPTO as an Edison Scholar.

181. U.S. PATENT & TRADEMARK OFFICE, PATENT PUBLIC ADVISORY COMMITTEE 2018 ANNUAL REPORT 16 (2018), *available at* https://www.uspto.gov/sites/default/files/documents/PPAC_2018_Annual_Report_1.pdf (supporting the Expanded Collaborative Search Pilots and IP5 Patent Cooperation Treaty Collaborative Search and Examination Pilot).

determining whether the underlying theory of change is actually sound.[182] Mechanism experiments support testing of the core underlying assumptions without requiring an extensive randomized controlled experiment;[183] as described in Part II, to the extent they simulate rather than actually effect policy changes, they impose fewer administrative burdens.

D. STEP 4 ("T"): SPECIFYING THE TEST STRATEGY

Rigorous piloting involves selecting a test strategy that is appropriate to the question being asked. A "mixed methods" strategy of getting feedback, e.g., through surveys, interviews, case studies,[184] and statistical (causal and non-causal) approaches is often best. Regardless of the specific method of evaluation, though, each of the other parts of specifying a pilot: articulating the question to be addressed, a theory of change, the evidence, how to measure it, and matching resources to the pilot, will require attention and in some cases, consultation with stakeholders, cost and convenience trade-offs, and implementation risk. In some cases, developing the right metric will be transformative, in others, framing the question correctly can unlock greater learning.

But when the question is, "did policy X have Y effect," experimental or quasi-experimental approaches are understood to yield the best information. Volumes have been written about appropriate methods of evaluations.[185] But also worth underscoring is what government pilots have revealed about the value of rigor, and the wasted opportunity associated with the lack thereof.

Section 1115 of the Social Security Act allows states that receive approvals from HHS to test and evaluate new approaches for delivering Medicaid services using funds that would not otherwise be eligible.[186] Taking advantage of these waivers, by 2016, nearly three-quarters of states implemented at least part of their programs under demonstrations.[187] To evaluate what had been learned, the Government Accountability Office ("GAO") reviewed demonstration projects in eight out of 15 states. The results were not good. The GAO concluded that a lack of care in how the evaluations were carried

182. *See* Jens Ludwig et al., *Mechanism Experiments and Policy Evaluations*, 25 J. ECON. PERSP. 17, 20–22 (2011).

183. *See id.* at 17–19.

184. *See, e.g.*, *Digital Health Software Precertification (Pre-Cert) Program*, FOOD & DRUG ADMIN., https://www.fda.gov/medicaldevices/digitalhealth/digitalhealthprecertprogram/default.htm (describing the FDA's selection of nine companies to participate in a pilot program to support streamlined approvals for software-based medical devices).

185. For some references, see website of the companion workshop to this paper, available at *Workshop on Rigorous Policy Pilots*, PENN LAW (May 30, 2019), http://law.upenn.edu/institutes/ppr/policypilots.

186. 42 U.S.C. § 1315(a) (2012).

187. U.S. GOV'T ACCOUNTABILITY OFFICE, MEDICAID DEMONSTRATIONS: EVALUATIONS YIELDED LIMITED RESULTS UNDERSCORING NEED FOR CHANGES TO FEDERAL POLICIES AND PROCEDURES 1 (Jan. 2018), *available at* https://www.gao.gov/assets/690/689506.pdf.

out, reported, and shared compromised their usefulness.[188] The lack of control groups in four out of the eight states, insufficient sample sizes and response rates for surveys, and the inability to isolate the impact of the Medicaid interventions from other changes limited what could be concluded about the effectiveness of demonstrations.[189] Rather than resource deficits being the problem—in 2015 alone, states spent over $100B, or one-third of Medicaid program expenditures on "demonstrations"—the problem seems to have been systemic attention deficits, to specifying upfront how evaluations could meaningfully be done,[190] for example by implementing them in a way that would support experimental or quasi-experimental evaluation, as well as, perhaps a lack of a bona fide intent to test. These "significant limitations . . . affected their usefulness in informing policy decisions." [191]

This description suggests that the lack of close alignment between those requesting and those implementing pilots can drastically reduce their value as a learning exercise.[192] One way to avoid this result is by ensuring that interventions are grounded, or at least framed, in terms of the lived experiences of those required to implement them. Presenting "rigorous pilots" as tools for advancing and testing assumptions and theories developed from the "bottom up" can distinguish them from "top-down" evaluation mandates that are not well integrated into agency decision-making.

The Medicare Section 1115 experience also highlights the importance of upfront design in setting out the conditions for experimentation. For example, specifying the control group in a quasi-experimental, after the fact, framework is often more difficult than in an experiment implemented from the start with randomization. For an evaluation carried out for the Center for Medicaid Services, researchers used a "difference-in-differences" framework[193] to evaluate the extent to which outcomes improved in states with home health agencies ("HHAs"). Though the treatments were not randomized, by comparing outcomes in the states did and did not receive the

188. *Id.* at 1.

189. *Id.* at 13.

190. *Id.* at 1.

191. *Id.* at i.

192. *See, e.g.*, Eileen M. Ahlin, *Conducting Randomized Controlled Trials with Offenders in an Administrative Setting*, 36 AM. J. OF EVALUATION 164, 168–72 (2015) (discussing, in the context of a series of RCTs carried out by a state agency, the resistance of Administrators and staff to pilots and how they were surmounted).

193. *See* ALYSSA POZNIAK ET AL., ARBOR RESEARCH COLLABORATIVE FOR HEALTH & L&M POLICY RESEARCH, EVALUATION OF THE HOME HEALTH VALUE-BASED PURCHASING (HHVBP) MODEL: FIRST ANNUAL REPORT 2, 24 (2018), *available at* https://innovation.cms.gov/Files/reports/hhvbp-first-annual-rpt.pdf. For an example of a differences-in-differences analysis carried out in the patent context, see Chien et al., PowerPoint Presentation, *Flight from Quantity . . . Flight to Quality? A Differences in Differences Analysis of Patent Applications and Complaints Following Patent Reform*, at slides 25–26 (Oct. 23–24, 2018), *available at* https://www.ssrn.com/abstract=3320907.

treatment, the researchers were able to draw some conclusions.[194] However, this required the use of a "hybrid comparison group strategy" to ensure that the two groups were appropriately matched.[195] The report "acknowledge[d] that the hybrid strategy . . . is complex. As part of our work for future reports, we will examine alternative approaches to simplify and refine our comparison group methodology."[196]

The choice of method will often be opportunistic and contextual. In a recent rulemaking procedure, the Department of Education blessed the use of randomized control studies, regression discontinuity design, and single-case design evaluation techniques in its award of grants in support of evidence-based educational policy-making.[197] The CFPB similarly has signaled its own use of within-subject and between-subject research methods, as well as survey, in-person, and other forms of evaluations.[198]

E. STEP 5 ("E"): EVIDENCE OR THERE'S NO EASY WAY TO MEASURE X

Specifying the criteria to use for evaluating a change in law or policy is one of the most crucial steps in designing a rigorous policy pilot.[199] Cary Coglianese has distinguished between the use of substantive and process outcomes for evaluating regulatory policy.[200] Substantively, metrics that

194. POZNIAK ET AL., *supra* note 193, at 2–3.

195. *See id.* at 2.

196. *Id.*

197. *See* Applications for New Awards; Education Innovation and Research (EIR) Program —Early-Phase Grants, 84 Fed. Reg. 1,093, 1,096 (Feb. 1, 2019) (defining a regression discontinuity design study as one that "assigns the project component being evaluated using a measured variable (e.g., assigning students reading below a cutoff score to tutoring or developmental education classes) and controls for that variable in the analysis of outcomes," and a single-case design study as one that "uses observations of a single case (e.g., a student eligible for a behavioral intervention) over time in the absence and presence of a controlled treatment manipulation to determine whether the outcome is systematically related to the treatment" (emphasis omitted)).

198. Johnson & Leary, *supra* note 48, at 187.

199. Take the example of e-commerce. When a company sends an email, it ultimately wants to support a transaction with the target recipient. However, a downstream purchase cannot easily be attributed back to a single email. Thus, email marketers generally rely on two other metrics to gauge the successfulness of an email campaign: open rate (was my email even opened?) and "click through rate" (did someone click a link within the email?). *See Email Marketing Benchmarks*, MAILCHIMP, https://mailchimp.com/resources/email-marketing-benchmarks (last updated Mar. 2018) (indicating that based on their analysis of email campaigns, "[t]he most opened emails are related to hobbies, with an open rate of 27.35%," followed by "[e]mails sent by government entities" (26.52% open rate), then the arts (26.03%)). Industry averages are around 21%. About 2.4% of opened emails are clicked through. These metrics are highly imperfect though because they fail to capture, for example, a scenario in which a person reads about a campaign through the email, then buys something from the firm at a later date or through a different medium, which is common. However, the metrics still provide objective "proxies" for the desired impact observable at scale.

200. Cary Coglianese, *Measuring Regulatory Performance*, OECD (Aug. 2012), https://www.oecd.org/gov/regulatory-policy/1_coglianese%20web.pdf.

implicate the effectiveness of a regulation in solving a problem, compliance with the regulation, and its cost-effectiveness are often going to represent important considerations. Cost and observability dimensions are likely to also influence the decision. "Precursors" (e.g., the cleanliness of a restaurant as a precursor to the outcome of foodborne illness) or "proxies" (e.g., hospital admissions by patients with relevant symptoms as a proxy for negative health effects) can provide attractive alternative measurement strategies when direct measures are costly or impractical to observe.[201] But only to the extent a strong causal connection to direct measures can be established.[202]

Even determining the appropriate "dimensionality" and unit of a metric can be hard. Take for example, the goal of the Department of Labor's Mine Safety and Health Administration ("MHSA") to reduce the number of miners dying in workplace accidents. Simply measuring this number per year would provide an inaccurate measure of effectiveness given variation year over year in the number of miners working in any given year, and also, potentially, the sensitivity of the measure to outlier events. Due to these factors, the MHSA has chosen to track a five-year rolling average of fatalities per 200,000 hours worked.[203]

Comparative process and outcome benchmarks, when available, can quantify that which is otherwise difficult to measure. The availability of "proxies" will depend heavily on the agency and circumstances. However, "synthetic" feedback can also be created where actual feedback takes too long to generate or is otherwise unavailable.[204] Depending on the treatment to be tested, human or artificial intelligence expert evaluators could be consulted. Whatever approach or combination of approaches is used, the methodology for evaluation should be transparent and clear, in order to build trust in its assessment.

F. STEP 6 ("R"): RESOURCES

Rigorously evaluating government policies requires human and financial resources. Carrying out a lab test requires paying participants. Not all agencies—at least at this point—have embedded data scientists or testing teams. One way to support rigorous evaluation is to fund it. The Department of Labor is not a scientific agency, but it has set aside up to 0.5% of its budget for evaluation and put its Chief Evaluation Officer into agency leadership.[205]

Evidence-focused initiatives within the federal government and non-profit groups that connect governments with individual academics that can

201. *Id.* at 32.

202. *Id.*

203. Harris, *supra* note 147, at 106–07.

204. In the patent system, for example, only 1–2% of patents are litigated at any point in a patent's 20-year lifetime, making organic court vetting a limited option.

205. *Primer: Strengthening a Learning Culture at the Department of Labor*, INDIAN HEALTH SERV., https://www.ihs.gov/dper/evaluation/evaluationresources/primer (last visited June 10, 2019).

also provide support. The Office of Evaluation Services within the General Services Administration supports agencies by carrying out rigorous tests of policy program implementation variants. From 2015 to early 2019, it has carried out over 60 rigorous tests that leverage behavioral science insights with over 18 agencies, in areas ranging from retirement security to economic opportunity.[206] Embedding researchers into agencies is another model that some have followed. The HHS' EITC work was supported in part by funding from the Ford Foundation, ostensibly to support the embedding of researchers into various parts of the initiative.[207]

Working directly with stakeholders and academics, by offering funding or the attention of regulators is another tactic. The CFPB's "Pitch a Pilot" program offers a way for stakeholders to suggest regulatory ideas for the agency to test, sparked by "notic[ing] something about a financial regulation that, if improved, could better foster consumer-friendly innovation."[208] Firms including American Express have accepted the Bureau's invitation.[209] In 2017, HUD published a notice in the Federal Register announcing that it was accepting unsolicited proposals for evaluation research, consistent with carrying out its mission.[210] The Notice specifically indicated that "HUD values demonstrations as a method for evaluating new policy and program initiatives and significantly advancing evidence-based policy, especially when *rigorous random-assignment methods are feasible*."[211] Further reducing the cost of evaluation, the Notice indicated, HUD had entered into interagency agreements with Census to link data from demonstrations with administrative data.[212] The success of these efforts remains to be seen but to the extent they ask open-ended, rather specific-specified questions, they can support a richer exchange of ideas and evidence.

IV. CONCLUSION

Through caselaw and examples, this Essay has demonstrated the legality, feasibility, and desirability of evolving law and policy using structured experiments. While it is informed by a recognition of the administrative state's comparative strengths, the Essay's purpose is to improve, through information, the development and implementation of law and policy. Though

206. *Work*, OFFICE OF EVALUATION SCIENCES, https://oes.gsa.gov/work (last visited June 10, 2019).

207. Gueron, *supra* note 166, at 11.

208. *Let's Collaborate: Pitch a Pilot*, CONSUMER FIN. PROTECTION BUREAU, *available at* https://www.consumerfinance.gov/about-us/innovation/pitch-pilot (last visited June 10, 2019).

209. CONSUMER FIN. PROT. BUREAU, INCREASING SAVING AT TAX TIME AND PROMISING PRACTICES FOR THE FIELD 1, 7–8, 16–27 (2015), *available at* https://files.consumerfinance.gov/f/201509_cfpb_increasing-saving-at-tax-time-and-promising-practices-for-the-field.pdf.

210. Authority to Accept Unsolicited Proposals for Research Partnerships, 82 Fed. Reg. 28,333, 28,335 (June 21, 2017).

211. *Id.* (emphasis added).

212. *Id.*

other accounts have discussed the idea of "test cases,"[213] this Essay explores the generative use of "tests" by administrative agencies to fill information gaps in law and policy. Many of the cautions and lessons around testing, including test hygiene, replicability, and disclosure of failures can also apply here and are saved for future explorations.

If the premise of this Essay—that using rigorous piloting is legal, feasible, and desirable—is correct, then it is incumbent upon policymakers to think about how to make it easier. The descriptions above and discussions at the companion workshop to this Essay reveal a number of suggestions for doing so: Clear obstacles to federal agency RCTs imposed, for example, by the Paperwork Reduction Act ("PRA"), reduce burdens associated with receiving feedback from stakeholders in general imposed through the Paperwork Reduction Act, create a networking mechanism for connecting federal agencies and academics interested in collaborating on rigorous policy pilots, provide dedicated capacity within government to support policy development through experiments (similar to what OES does for program implementation), make it easier for agencies to get access to the resources they need to carry out pilots, create safe harbors for agencies around the disclosure of procedurally adequate experimentation and its results, support evaluation and piloting mandates within statutes.

Academics and advocates interested in having a policy impact and their evaluators—peers, editors, and tenure review committees—also have much to contribute. Currently, journal articles typically contain recommendations for action in terms of new or revised laws or policies. But just as products aren't developed overnight, it's unrealistic to expect new laws or policies to be introduced without first testing and vetting them. This insight is not new. The recognition in biomedical innovation a decade ago of a "chasm" between medical research and patient needs has been a fillip to the field of translational science.[214] But law and policy advocates and scholars also can help bridge the "valley of death" between early stage ideas and enacted laws, rules and practices by proposing, and where feasible, implementing lab versions of thoughtful policy pilots.

Each of these steps is non-trivial, requiring a deep understanding of the powers and authorities of the relevant agency, the context in which it operates, the content and footprint of high priority questions, and the data sources from which meaningful metrics can be developed. Though the implications or recommendations section of a law review article or report often get the least amount of space, designing and proposing a pilot often requires novel contributions that support not a fixed policy "product," but a dynamic policy "process" for testing and refining a policy idea before it is fully

213. Arti Rai, *Patent Validity Across the Executive Branch*, 61 DUKE L.J. 1237, 1270 (2011).

214. Declan Butler, *Translational Research: Crossing the Valley of Death*, NATURE (June 11, 2008), https://www.nature.com/news/2008/080611/full/453840a.html.

implemented. A broader recognition of the importance and originality of this "translational work"—by law review and other editors—would shift the focus of academic writing from "admiring the problem"[215] to "solving the problem."

This Essay has demonstrated that rigorous policy pilots are legal, feasible, and desirable. It has provided a framework for proposing a rigorous pilot that reflects the accumulated wisdom of agencies and practitioners that have worked for decades to test, learn and develop interventions that can help the American public, "MATTER" (questions that **m**atter, relevant **a**uthority, the underlying **t**heory of change, **t**esting protocol, **e**vidence and **r**esources), and, in an online appendix shows how it can be applied. Why not give it a try?

215. As to a problem, spending a disproportionate amount of time to "understand it fully, we know the impacts, and the secondary impacts" and not in deciding solutions to the problem as described in Eric Holdeman, *Admiring the Problem*, GOV'T TECH. (July 3, 2017), http://www.govtech.com/em/emergency-blogs/disaster-zone/admiring-the-problem.html.