Policy-makers and evidence: double standards? A field experiment (Pre-Analysis Plan)*

Angele Delevoye† and Trevor Incerti‡
Yale University

April 29, 2019

Abstract

A recent, bipartisan concern in Congress is the need for more evidence-based policy-making. We propose a field experiment's idea to test whether legislators are capable of rising up to their own intentions and goals. Are policy-makers really more receptive to higher causal evidence standards when receiving policy-information? And even if they want to be, do they have sufficient knowledge of evidence standards to be able to differentiate between different levels of information? This pre-analysis plan lays out initial details for an experimental design dedicated to answering these questions. We include a power analysis based on these still preliminary details.

*

†angele.delevoye@yale.edu
‡trevor.incerti@yale.edu

1 Introduction

The need for more evidence-based policy-making is the object of a rare bipartisan agreement. Congress passed the Foundations for Evidence-Based Policymaking Act in 2018, a legislation sponsored by then-House Speaker Paul Ryan (R-WI) and Senator Patty Murray (D-WA). Only 17 representatives voted against, and it passed the Senate by Unanimous consent. Policy-makers' intentions are clear, but do their behaviors and knowledge align with those intentions? Are policy-makers really more receptive to higher causal evidence standards when receiving policy-information? And even if they want to be, do they have sufficient knowledge of evidence standards to be able to differentiate between different levels of information? In order to investigate these questions, we develop an idea for a field experiment. We would reach out to policy-makers with new information on a policy, and would randomize the source of this new information (more concretely, the kind of study from which the information came from and its level of scientific evidence: from a simple regression to a well-conducted RCT). We would also randomize whether we provide information to the policy-maker as to what the scientific evidence means (telling them for example that a RCT is considered the golden standard in causal inference and evaluation studies).

This pre-analysis presents the theories and existing literatures we would rely on and hope to contribute to. We are especially hoping to build on the few similar experimental studies that have reached out to policy-makers in American Politics. Then we get into the details of our proposed experimental design: treatment (choice of policy and evidence standard), logistics (partnership with a 3rd party organization), outcome measures and ethical considerations. These details on the design are still preliminary, and several open questions remain. We finish this pre-analysis plan with a power analysis for the experiment, based on the details that are set at this stage.

2 Theory

The first literature of interest is the growing literature on the use of evidence and expertise in policy-making: whether, how and when is expertise used by policy-makers. The general wisdom emanating from the literature on Congress has long been that evidence and information do not matter. Legislators are focused on reelection (Mayhew 1974), power within Congress (Fenno 1973), and answering to local pressures and issues (Fenno 2002).Schick (1976) concluded in 1976 that "Congress is not a natural habit for policy analysis". Given those institutional constraints and lack of time and resources, the best Congress can do is "muddle through" (Lindblom 2018). But more recently, scholars have called for a more nuanced understanding of the use of evidence in policy-making Patashnik and Peck (2016), despite well-identified methodological challenges (Mandell & Sauter 1984). Multiple case studies have shown that evidence is used in policy-making, but we need a better understanding of how (much), when and why. Scholars have also looked at the supply side of scientific evidence, and have looked at what scientists can do to maximize the use of scientific evidence. Cairney (2016) lays out the two types of shortcuts used by legislators, a rational one and a more irrational form of decision-making drawing on emotions, beliefs and habits. He argues that scientists have mostly focused on the former, and that they should do a better job at accompanying evidence with simple stories to exploit the emotional or ideological biases of policymakers. We are hoping to test and measure legislators' ability to respond to the first, rational type of stimulus.

Depending on what and who the source of the information ends up being in the experiment, we could also look at literatures on messenger effects (does the source of information matter?), relations between diverse groups and policy-makers, etc. The initial goal is to conduct this experiment with the U.S. Congress, but we might end up conducting it at the State or more local levels (cf. below). In that case, literatures on state and local politics would come into play. Finally, this project might address broader, more theoretical questions

on the functioning of our current democracies, and the potential gaps between reality and the ideals of deliberative democracies. We are also hoping that this project will contribute to ongoing philosophy of science discussions. What is the value of our different methods outside of the research community bubble, how are they perceived and understood by different targets (policy-makers for this specific experiment, but a similar experiment could later be conducted with citizens)?

In American Politics, researchers reaching out to legislators under a randomized experimental design has been rare. Table 1 below summarizes the experiments we are hoping to draw inspiration from. Our design will probably end up being a mix of Kalla and Broockman (2016) and Butler and Broockman (2011)'s experiments. Our factorial design will come close to Butler and Broockman (2011)'s 2x3 design (and we might need to use their contact list for state legislatures). We are also hoping to emulate Kalla and Broockman (2016)'s use of a 3rd party organization to reach out to the legislators, as well as their more nuanced outcome measure: both rate of response to emails and the seniority of the person who agreed to meet with the organization. We believe that roll call votes as the outcome measure, used by Bergan (2009), is only possible for an experiment extremely targeted on one bill and one Chamber.

To our knowledge, no similar experiment has tried to reach out to bureaucrats. There are very good, obvious reasons for this: constituents and groups reaching out is business as usual for a legislative office, and part of a legislator's job (and incentives) is to listen to constituents and groups. Reaching out to bureaucrats would be much less natural, we would need a legitimate reason and credible source (bureaucrats would probably not answer to advocacy groups or think tanks - they might to more research-oriented and institutional sources). Very aware of these challenges, we do not start this project with bureaucrats in mind. But we are also convinced of the substantive interest such an experiment would have. Therefore we will remain open to the possibility of coming back to this topic, should the

opportunity arise.

Table 1: Audit experiments conducted with U.S. policy-makers

Reference	Federal/State	Treatment #	Treatment	Design	Outcome	3rd party?
Bergan (2009)	House of Representatives of New Hampshire	1	Getting contacted by activists	Matched pairs (multimembers districts) Randomization within party and district stratas	Roll Call votes	Yes: coalition of public health-related groups organizing a grassroots email lobbying campaign by activists
Butler and Broockman (2011)	4,859 state legislators (44 states)	2x3	Black or white name and party (D/R/blank) of email sender	Block randomization by state, chamber, party, and whether legislator is up for reelection	Rate of response to emails	No
Kalla and Broockman (2016)	US Congress (191 offices) These were the offices that had not sponsored the bill yet	1	Reveal in email that prospective attendees had contributed to campaigns	Blocks of 3 offices: closest similarity on multiple covariates 1 treated, 2 control in each of the 64 blocks	Rate of response to emails and seniority of proposed meeting	Yes: liberal political organization trying to set meetings between offices and constituents who had previously given to campaigns. Goal of the meetings: rally support for a bill banning a chemical.
Doberstein (2017)	1,108 Canadian bureaucrats	2x2	Source of the policy information (academic, think tanks, research-based advocacy groups)	Sources in treatment groups were falsified Pre treatment survey for covariates	Credibility assessment of each of 5 research articles Based on summaries And ranking of the 5 articles	No.
Zelizer (2018)	18 bills 76 state legislators	1	Assigned to in-person briefings by a committee staffer	Treatment assigned at legislator-bill dyad level block RA	Cosponsorship of bills Roll-call votes	No.
Butler, Nickerson, et al. (2011)	New Mexico State House 70 legislators	1	Received district-specific survey results on constituents' opinions on a bill	35 matched pairs	Cosponsorship of bills Roll-call votes	Use of University Logo And email from own researchers' address
Butler, Karpowitz, and Pope (2012)	489 legislatives offices 23 states 1,036 letters	2x4	Received letter ethnicity of sender varied 2 ethnicities type of letter varied 4 types (policy vs service, level of knowledge)	35 matched pairs	Rate of response Roll-call votes	Yes: actual individuals 200 students at BYU Opened post office boxes in their hometown

3 Design

3.1 Treatment: a factorial design

We would reach out to policy-makers and provide them with information on a new policy finding coming from a study. We would have a 2x4 factorial design with two treatments: (i) the evidence standard used in the study (low, middle, high, qualitative) and (ii) whether we add a paragraph explaining the evidence standards and what they mean. Table 2 synthesizes our treatment arms. We give more details about the type of evidence standards and policies we would use in the following paragraphs.

Table 2: 2x4 factorial design

	Lower Tier	Middle Tier	Higher Tier	Qual studies
No information	Control	ATE 1	ATE 2	ATE 3
Information	ATE 4	ATE 5	ATE 6	ATE 7

3.2 Evidence standards

For evidence standards and their descriptions, we can rely on evidence standards already set in some laws and used by the Department of Education (DoE) and Department of Labor (DoL). The DoL has three tiers of evidence: high causal evidence (mostly RCTs), moderate causal evidence (strong non-experimental designs or RCTs with high attrition) or low causal evidence studies. The DoE has similar standards (cf. table Appendix A). The upside of using these standards is obvious: they are used in real life policy-making, the definitions exist and are clearly specified in laws and implementation texts, and they are associated with specific projects and studies (see below). One downside is that they probably mean restricting the experiment to quantitative studies [1]. We will need to think about whether we want to add qualitative studies as an additional category. We will also need to check

[1] We have not yet looked at the databased and associated studies. We doubt it, but it is possible that qualitative studies are already incorporated into the databases and one of the 3 evidence standards

whether qualitative studies are used in the DoE and DoL databases, and if yes, under which tier of evidence (this would reduce our design to a 2x3).

3.3 Choice of policy

The policy and specific findings we will report on are yet to be determined. Ideally, we would use similar findings coming from 3 or 4 studies belonging to different tiers of evidence. The DoE and DoL each have comprehensive databases of most studies conducted in their respective policy areas, their findings and the evidence tier to which they belong. We can leverage these existing databases (What Works Clearing House for the DoE, CLEAR for the DoL) in order to find studies we can use in our experiment.

We would work with the IRB to determine what we can or cannot do. We will probably need to find 3 or 4 studies that reach similar results with different methods (and if the results are not perfectly similar, they need to be close enough for us to be able to use the same one-sentence presentation of them in our emails without any deception on our part). Depending on the studies that we find, we might need to be more or less specific in the formulation of the results in our email (it could be anything from a very broad result such as 'small classes are better' to more specific results such as 'adding curriculum x to classes y raises z'). The results would also ideally need to be somewhat surprising, or at least new, in order to maximize the level of new information we are bringing, and raise the interest in and incentives of meeting with us.

3.4 Outcomes

What outcomes we would measure and how is still an open question. The more realistic solution would be to measure response rates to our email. If we are more ambitious and partner with a 3rd party organization, we could ask to set meetings and use the seniority of whoever we get a meeting with as another measure of the importance given to our approach. The model for this would be Kalla and Broockman (2016). If we actually set meetings (through

a third-party organization), more qualitative observations of policy-makers? understanding of and responsiveness to evidence standards could be gathered.

3.5 Partnership with 3rd party organization

The source of the email will be important, and has yet to be determined. In order to maximize the realism and reach of the experiment, we will probably need to partner with a third-party organization. Emails will come from this organization, not from us ? once again, modelling Kalla and Broockman (2016)'s experiment. We could reach out to think tanks, policy labs, advocacy groups to partner with for this project. If we work with state and local policy makers, we will try to find a group with a large geographic and regional presence. It would be better to reach out to policy-makers in Idaho with a group with some presence if not in Idaho, at least in the northwest.

3.6 Federal or state level?

Conducting the experiment at the federal level would be substantially interesting (we are using recent initiatives happening at the federal level as motivations for this project). It might also be logistically easier to organise and attend meetings. But we might end up conducting it at the State or more local levels, for multiple reasons. First, we might encounter power issues if we have multiple treatment arms and a fixed N of 535. Second, only two policy areas have existing, well-defined evidence standards and associated studies (cf. below): education and labor policies. Depending on which specific policy we end up choosing, it might make substantially more sense to conduct the experiment at the local and state level. Third, it might be more realistic and easier to set meetings with legislators at the local level. Finally, there is a substantive interest as to whether state and local politics can help compensate for the shortcomings of the federal level, both in terms of quantity (gridlock) and quality (less polarized, more time to act).

4 Ethical considerations

5 Power Analysis

6 Conclusion

References

Bergan, D. E. (2009). Does grassroots lobbying work? a field experiment measuring the effects of an e-mail lobbying campaign on legislative behavior. *American politics research*, *37*(2), 327–352.

Butler, D. M., & Broockman, D. E. (2011). Do politicians racially discriminate against constituents? a field experiment on state legislators. *American Journal of Political Science*, *55*(3), 463–477.

Butler, D. M., Karpowitz, C. F., & Pope, J. C. (2012). A field experiment on legislators? home styles: service versus policy. *The Journal of Politics*, *74*(2), 474–486.

Butler, D. M., Nickerson, D. W., et al. (2011). Can learning constituency opinion affect how legislators vote? results from a field experiment. *Quarterly Journal of Political Science*, *6*(1), 55–83.

Cairney, P. (2016). *The politics of evidence-based policy making*. Springer.

Doberstein, C. (2017). Whom do bureaucrats believe? a randomized controlled experiment testing perceptions of credibility of policy research. *Policy Studies Journal*, *45*(2), 384–405.

Fenno, R. F. (1973). *Congressmen in committees*. Little, Brown.

Fenno, R. F. (2002). *Home style: House members in their districts (longman classics series)*. Longman Publishing Group London, England.

Kalla, J. L., & Broockman, D. E. (2016). Campaign contributions facilitate access to congressional officials: A randomized field experiment. *American Journal of Political Science*, *60*(3), 545–558.

Lindblom, C. (2018). The science of ?muddling through? In *Classic readings in urban planning* (pp. 31–40). Routledge.

Mandell, M. B., & Sauter, V. L. (1984). Approaches to the study of information utilization in public agencies: Problems and pitfalls. *Knowledge*, *6*(2), 145–164.

Mayhew, D. R. (1974). *Congress: The electoral connection* (Vol. 26). Yale University Press.

Patashnik, E. M., & Peck, J. (2016). Can congress do policy analysis? *Governing in a Polarized Age: Elections, Parties, and Political Representation in America*, 267.

Schick, A. (1976). The supply and demand for analysis on capitol hill. *Policy Analysis*, 215–234.

Zelizer, A. (2018). How responsive are legislators to policy information? evidence from a field experiment in a state legislature. *Legislative Studies Quarterly*, *43*(4), 595–618.

7 Appendix A: Department of Education's Evidence tiers