Legislators and use of evidence: double standards? A field experiment (Pre-Analysis Plan)*

Angèle Delevoye† and Trevor Incerti‡
Yale University

May 28, 2019

Abstract

The bipartisan Foundations for Evidence-Based Policymaking Act of 2018 stresses the need for evidence-based policy-making. We propose a field experiment that tests whether legislators are actually more receptive to higher standards of research evidence when receiving and evaluating policy information. A related question is if legislators possess sufficient knowledge of evidence standards to be able to differentiate between varying standards of research quality. This pre-analysis plan proposes a field-experimental research design dedicated to answering these questions, and pre-registers the procedures that will be used to conduct the analysis.

*

†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 - 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 standards of evidence when evaluating policy-information? Moreover, even if policymakers want to adopt policies based on high-quality evidence, do they have sufficient knowledge of evidence standards to be able to differentiate between research of varying quality?

We propose a field experiment to test: (1) whether policymakers give more credence to high quality research, and (2) if policymakers can recognize differences in research quality. Our research design would entail contacting policy-makers with information regarding research on a potential educational policy while randomizing the kind of research design from which study derives its evidence - from a simple regression to a well-conducted RCT. We would also randomize whether we provide the policy-maker with information about varying standards of evidence within academic research (e.g. informing them that an RCT is considered the gold standard in causal inference and policy evaluation studies).

Our pre-analysis first presents the theory and existing literature we hope to advance. In particular, we hope to build on existing experimental studies that have contacted American policy-makers. Next, we present 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 conclude with a power analysis for the proposed experiment.

2 Theory

2.1 Science, research and democracy

This project may speak to broader theoretical questions on the functioning of our current democracies, and the potential gaps between reality and the ideals of deliberative democracies. In this sense, we hope this project will contribute to ongoing discussions within philosophy of science, such as perceptions of research methods and consumption of scientific evidence outside of the academy.

[Tensions and relationships between science and democracy]

2.2 Evidence in policymaking

A growing literature examines if, when, and how evidence and expertise is used in policymaking. The conventional wisdom from the this literature in the US Congressional context has long been that evidence and information do not matter. Legislators are focused on reelection (**?**), power within Congress (**?**), and responding to pressures from local constituents (**?**). **?** concluded in 1976 that "Congress is not a natural habit for policy analysis," and **?** states that the best Congress can do is "muddle through" given institutional constraints and lack of time and resources.

Recently, however, scholars have called for a more nuanced understanding of the use of evidence in policy-making (**?**), despite well-identified methodological challenges (**?**). Case studies have suggested that evidence is used in policy-making, but we need a better understanding of how (much), when and why. Scholars have also examined the supply of scientific evidence, and have looked how scientists can maximize the use of scientific evidence. **?** lays out two types of shortcuts used by legislators, rational decision-making and irrational 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

hope to test and measure legislators' ability to respond to the first, rational type of stimulus.

Experimental research on messenger effects (i.e. does the source of information matter?) may also be relevant to our research design. Source credibility theory hypothesizes that high-credibility sources are more effective in matters of persuasion than low-credibility sources and distinguishes between expertise and trustworthiness as sources of credibility (**?**). Additional research has found that trustworthiness (i.e. being "liked") is more effective than expertise (i.e. being "right") (**?**), implying that the source of evidence is important in addition to evidence quality. This implies that the institution responsible for the research may be highly important, and should therefore be held constant across treatment groups and should be a trusted, non-partisan institution.

Our initial goal is to conduct this experiment with the U.S. Congress, however it may be necessary to change our focus to the the state or local level for power or logistical reasons. In this case, we would need to consider literature on policy deliberation in state and local politics.

2.3 Legislator contact experiments

Randomized contact of legislators in American politics research remains relatively rare. **??** below summarizes previous field-experimental studies that have contacted legislators. Our design most closely resembles a mix of the experiments conducted by **?** and **?**. Our factorial design will come close to **?**'s 2x3 design (and may use their contact list for state legislatures). We also plan to emulate **?**'s use of a third party organization to contact legislators, as well as their more nuanced outcome measure: both rate of response to emails and the seniority of the person who has agreed to meet with the organization. The use of roll call votes as in **?** is likely only possible for an experiment that targets one bill in one Chamber.

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

Reference	Federal/State	Arms	Treatment	Design	Outcome	3rd party?
?	State (New Hampshire)	1	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
?	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
?	US Congress 191 offices that had not yet sponsored bill	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.
?	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.
?	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.
?	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
?	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 Experimental design

3.1 Treatment groups and randomization

Policy-makers will be contacted and provided with information on findings from a policy study. This will take the form of a 2x2 factorial design with two treatments: (i) the evidence standard used in the study (low vs high) and (ii) whether we provide policy-makers with information explaining the evidence standards (no explanatory information or explanatory information). **??** synthesizes our treatment arms. Additional details regarding the types of evidence standards and policies included in each arm are described in **??** below.

Given the small sample size inherent in any study using a sample of legislators, we will conduct block random assignment based upon a vector of pre-treatment covariates in order to increase the precision of our treatment effect estimates. Examples of pre-treatment covariates that can be used to create the blocks would be party, chamber, education level of the legislator (if data is available), closeness of the district, state, etc. In other words, legislators will be divided into blocks based on these pre-treatment covariates, then will be randomly assigned to a treatment condition within each block. Which treatment arm (i.e. form of contact) a policy-maker receives will be randomized using complete random assignment, and will in practice be conducted using the "complete_ra" function in the R package *randomizr* (part of the *DeclareDesign* suite).

Table 2: Treatment arms: 2x2 factorial design

	Lower Tier	Higher Tier
No information	Control	High and no info
Information	Low and info	High and info

3.2 Evidence standards

We rely on evidence standards and descriptions already adopted in some laws (Elementary and Secondary Education Act - ESEA 65, No Child Left Behind - NCLB 01, Every Student Succeeds Act - ESSA 15) and used by the Department of Education (DoE) and other federal agencies. Since 2015, the Every Student Succeeds Act encourages state educational agencies (SEAs), local educational agencies (LEAs), and schools to prioritize and include evidence-based interventions, strategies, or approaches. Funding for some program is conditional on proving that the program has been showed to work by a rigorous, evidence-based study. ESSA's definition of "evidence-based" includes 4 levels of evidence:

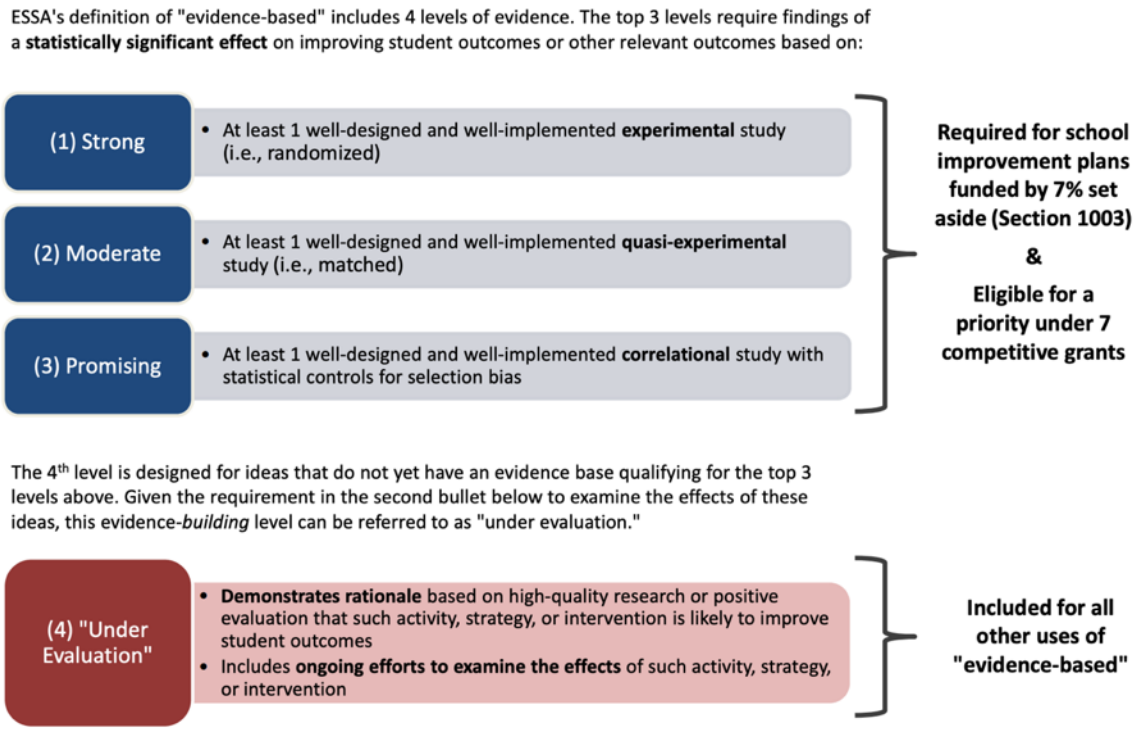


Figure 1: ESSA's tiers of evidence (Source: bill text and Results for America)

The DoE has extremely detailed implementation standards for ESSA, with very specific decision-trees to decide in which of their three quality tiers of evidence each study falls under. RCTs with low attrition, or RDDs satisfying the usual assumptions (integrity of the forcing variable, sample attribution, continuity, bandwidth and functional form choices) fall under

the high causal evidence category. The DoE also provide clear guidelines on how to deal with missingness, clusters, CACE/IVs designs (see **??** for detailed decision-trees used by the DoE).

Other federal agencies have followed the DoE's steps and developed their own standards (see **??**). For example, the DoL defines three tiers of evidence: high causal evidence (primarily RCTs), moderate causal evidence (strong non-experimental designs or RCTs with high attrition) or low causal evidence studies. The DOL defines the low causal studies as studies that do not meet criteria for a high or moderate evidence rating. Those studies, according to the DoL, show little evidence that the effects estimated in the study are attributable to the intervention being examined, and other factors are likely to have contributed to the results. They should be interpreted with caution.

We use these standards because they are ostensibly used in actual policy-making decisions. These definitions already exist, are specified in law and implementation texts, and have been referenced in previous projects and studies (see below). In order to maximize statistical power and increase the clarity of our research design, we will focus only the studies from the highest tier of evidence and those from the lowest tier. This will keep our experimental design to a 2x2 factorial design, as opposed to a higher dimensional design that would result from inclusion of all evidence tiers. One potential downside of this approach is that it may require us to restrict ourselves to quantitative studies.

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 2 studies belonging to different tiers of evidence. The DoE has a comprehensive database of most studies conducted in education policy (What Works ClearingHouse), their findings and the evidence tier to which they belong. In order to identify the studies which belong in the dataset, the Institute of Education Sciences (IES) within the DoE conduct literature scans, receives public nominations or suggestions from education

associations or other organizations, as well as input from state and federal policymakers.

We can leverage this existing database, or similar ones constructed by other federal agencies (see **??** for a more comprehensive list of federal clearinghouses), in order to find studies we can use in our experiment.

For ethical reasons we will need to locate studies that reach similar results using different methods (ethics are considered in more detail in the **??** section below). More specifically, the results of the studies will need to be similar enough to use the same one-sentence summaries in our messages without resorting to deception. 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 class size increase educational outcomes" to more specific results such as "adding curriculum x to classes y raises z"). Ideally, the results would also be surprising or new, in order to both maximize the level of new information we are providing, as well as to raise the incentives of meeting with an organization regarding the results. For example, providing policymakers with information regarding an educational policy that has already been adopted would be ignored for reasons unrelated to the content of the randomized treatments.

3.4 Outcomes

What outcomes we would measure and how is still an open question. The simplest outcome to measure would be email response rates. If we are able partner with a 3rd party organization, we will examine two outcome measures: (1) whether or not a meeting was established, and (2) the seniority of the individual with whom a successful meeting was granted. This design mirrors that of **?**. Setting meetings would also allow us to gather additional qualitative information analyzing the substance of discussions between the third party organization and policymakers, in addition to the content of the email responses .

3.5 Partnership with 3rd party organization

The source of the email has yet to be determined. In order to maximize the realism and reach of the experiment, we will ideally partner with a third-party organization seeking to establish meetings with policymakers. Emails will be sent from this organization, rather than from us. Once again, this design mirrors that of **?**. To this end, we will contact think tanks, policy labs, and advocacy groups as potential partners.

3.6 Email wording

This is an example of the types of emails we sent.

3.7 Federal or state level?

Conducting the experiment at the federal level would be of primary substantive interesting, as recent federal initiatives provide motivation for this study. In addition, it may be logistically easier to organize and attend meetings at the federal level, as these meetings could be centralized in Washington DC. However, we may be required to conduct the study at the state or local level for two primary reasons. First, we may encounter power issues with multiple treatment arms and a fixed N of 535 (see **??**). Second, only two policy areas have existing, well-defined evidence standards and associated studies: education and labor policies. If the policy we choose is primarily a topic of state and/or local level debate, it may be better to conduct the experiment at this level.

4 Ethical considerations

We recognize that reaching out to policy-makers with information on a policy can possibly involve ethical challenges. **?** lay out three types of ethical risks associated with experiments on public officials. The first is the use of deception: policymakers will be unaware that the emails we send are part of an experiment. However, the information we provide to policymakers will be real, and we will offer to organize actual meetings with an organization

that wishes to meet with policymakers. This implies that the only deception involved is the existence of experimental manipulation itself.

A second potential ethical risk is political harm to the legislator. To this end, the anonymity all legislators and their responses will be preserved: no data will be available that can connect an answer to a specific legislator - only averages will be made public.

A third risk is the potential burden our experiment would place on legislators' time. We will therefore do our best to keep our messages as short and simple as possible, while still containing all of the necessary information. We recognize that a meeting will take time out of legislators' already busy schedules. However, we believe that an office that agrees to the meeting would be interested in learning about the given policy, and indeed would be performing one of the normal function of policymakers - meeting with interest groups as a form of knowledge acquisition. This time investment would therefore not be wasted and would bring value to both the legislative office and the interest group, outside of and beyond the value of the experiment to us.

5 Experimental analysis

5.1 Estimation of treatment effects

As noted above, we intend to use block random assignment in order to increase the precision of our treatment effect estimates as well as to facilitate (preregistered) estimation of heterogenous treatment effects. Treatment effect estimates will therefore be calculated as the difference-in-means of the response rate from subjects in each of treatment groups and the response rate from subjects in the control group in each block, weighted by the number of subjects in each block relative to the total number of subjects. More formally, average treatment effects will be estimated as:

$$ATE = \sum_{j=1}^{J} \frac{N_j}{N} ATE_j$$

where J is the number of blocks, blocks are indexed by j, and $\frac{N_j}{N}$ represents the share of

subjects who belong to block j.

In practice, these differences-in-means will be calculated using a weighted least squares regression of response rate on treatment assignment, with weights corresponding to the inverse probability of treatment for each unit (IPW). All p-values will be calculated using randomization inference. As the reference/control group in the experiment is in effect the "No information / lower tier" treatment group, all effects must be interpreted in relation to this treatment. In other words, treatment effects should be interpreted as the change in response rate relative to the "No information / lower tier" group.

5.2 Heterogenous treatment effects

In addition, we will test for heterogenous treatment effects on legislator party and educational attainment. Our test will take the form of regressing our outcome variable on treatments conditional upon the data representing the covariate of interest. In other words, we will split our sample by subject attributes (party and educational attainment), then estimate conditional average treatment effects (CATEs) separately for each of these attributes. For example, to test for heterogenous effects by party, we will regress outcomes on treatments for all legislators in each party. Following estimation of CATEs, we will use randomization inference to test the null hypothesis that CATEs in both groups (e.g. Democrat and Republican) equal to the ATE.

We recognize that the search for heterogenous treatment effects often suffers from the multiple comparisons problem. In a dataset with a large number of covariates, if a large enough number of subgroups is examined, it is highly likely that statistically significant interaction effects will emerge merely by chance. In other words, the more significance tests are performed, the higher the likelihood of falsely rejecting the null hypothesis at least once. We minimize this risk be preregistering our heterogenous effects of interests, as well as by performing multiple comparisons corrections using Bonferroni, Holm-Bonferroni, and Benjamin-Hochburg adjustment procedures.

6 Power Analysis

We perform power analyses in order to estimate the maximum number of treatment groups that may be advisable to include in the experiment given sample size and budgetary constraints. Preliminary power analyses are depicted below, and derive their assumptions regarding hypothetical response rates and treatment effect sizes from past experiments. These preliminary power analyses are conservative estimates, as we did not include the effects of our future blocking strategy.

The first power analysis assumes a sample size of 535 respondents (i.e. the number of representatives and senators in the US Congress). In **?**, congressional offices granted meetings in response to requests roughly 50% of the time. However, as the organizations requesting meetings were political donors (whether revealed or not), we believe that such a success rate may be unreasonably high for our own experiment, and therefore conduct our power analysis assuming a more conservative 30% of offices will grant meetings. 30% therefore represents the assumed rate of successful meetings in the first treatment group (low information + no information provision), which may be thought of as the control group. In Treatment 2 - low quality evidence + information provision - we assume an average treatment effect of -10%. In Treatment 3 - high quality evidence + no information provision - we assume an average treatment effect of 5%. In Treatment 4 - high quality evidence + information provision - we assume an average treatment effect of 12.5%. For all treatment groups, we assume a standard deviation of 0.08. A power analysis using 10,000 simulations results in power of 0.86, 0.34, and 0.97 in treatment groups 2, 3, and 4, respectively.[1] A graphical depiction of this power analysis can be found in **??**.

We next conduct a power analysis using state legislatures. This theoretically could increase our possible sample size drastically, to roughly 5,000 respondents. However, we recognize that it would be logistically impractical for an organization to set up meetings with with

[1] Rounded to the nearest 100th decimal place.

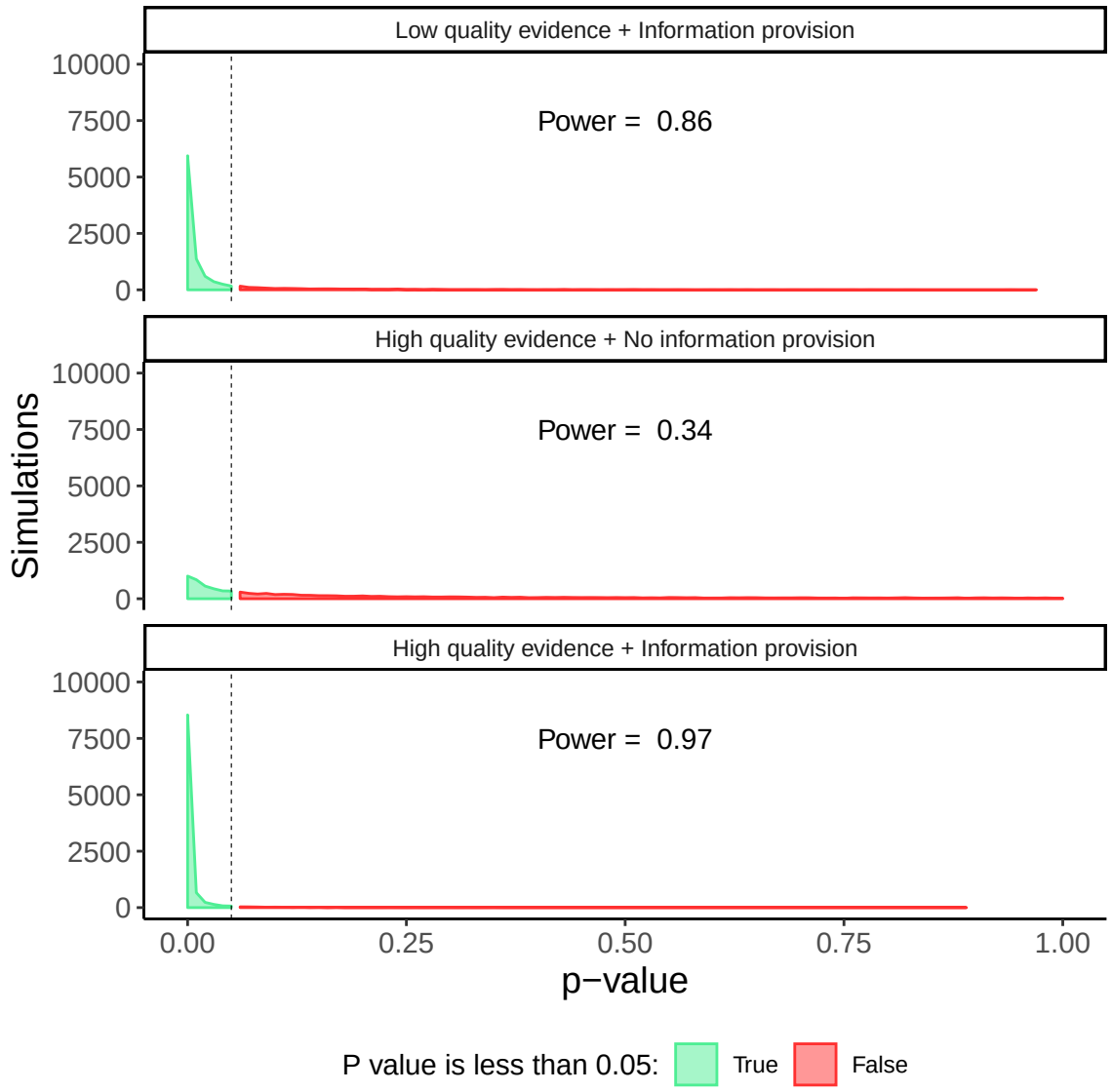


Figure 2: Power analysis: federal congress

such a large number of offices across such a wide geographical range. We therefore restrict ourself to a sample size of 1000, which could be achieved by sampling only from states in the Northeast United States, or from states where the organization has a local presence. Using the same assumptions of treatment effect sizes and variable that informed the power analysis in **??**, **??** depicts the statistical power results from this increased sample size. The larger sample size results in statistical power of 0.99, 0.55, and 1 in treatment groups 2, 3, and 4,

respectively.[2]

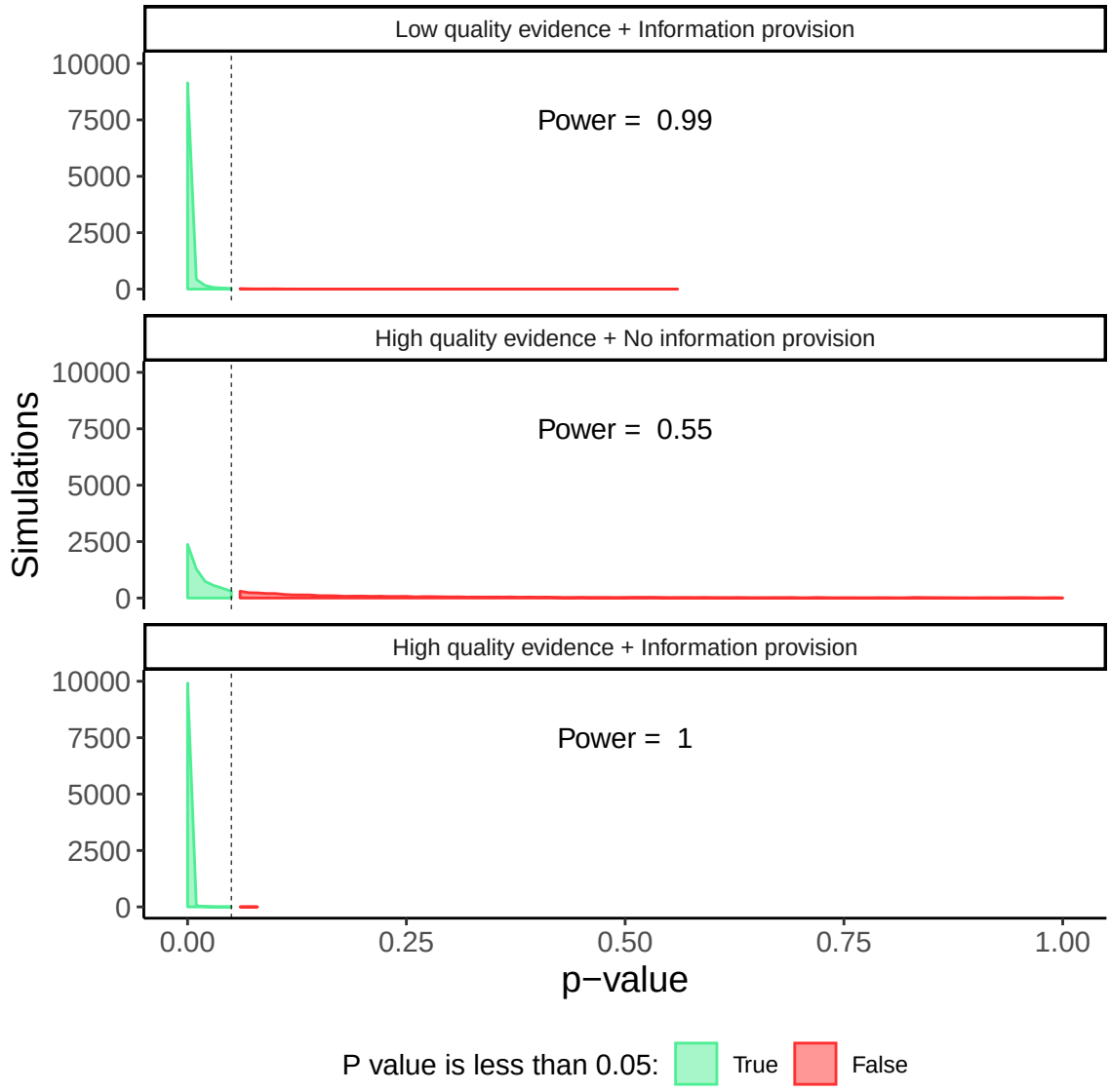


Figure 3: Power analysis: state legislatures

7 Conclusion

Our next steps on this project include finalization of some of the research design decisions that remain open (i.e. federal or state level subjects, partnership with an organization, and choice of the policy). We hope to finalize these details over the summer and to preregister a final

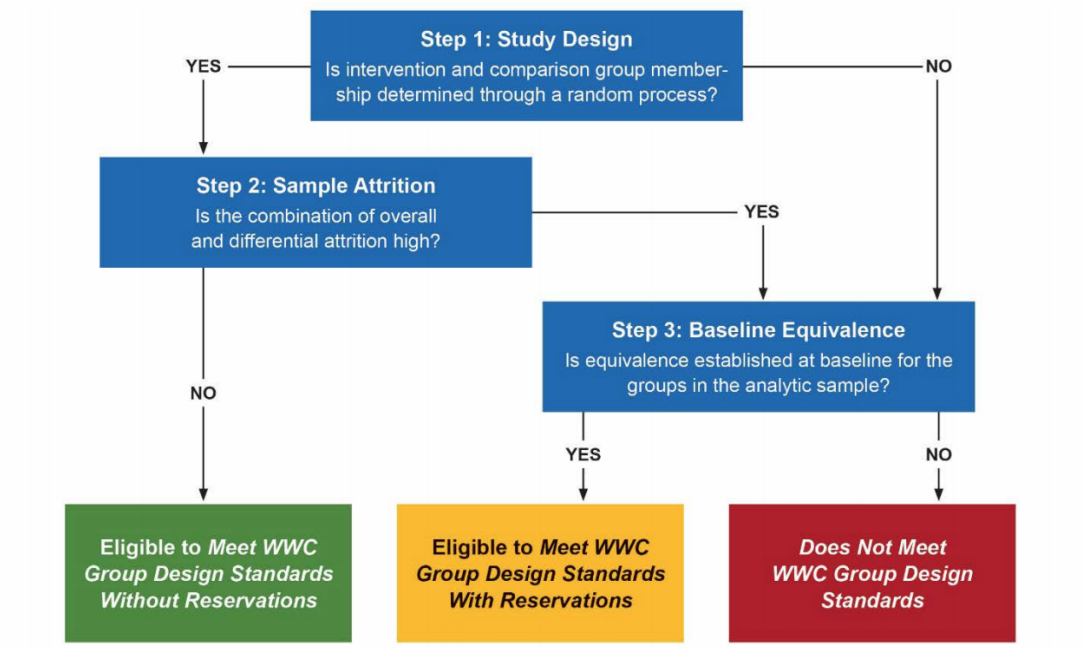
[2]Rounded to the nearest 100th decimal place.

Pre-Analysis-Plan with EGAP by the end of 2019. We would then conduct the experiment in 2020. We believe that this timeline would be ideal because of the political and electoral contexts expected in 2020. An election year increases the likelihood that parties will be seeking innovative policy ideas (and decreases the time spent on other legislative activities), and it is therefore possible that legislators will be particularly responsive to information on such policies.

A Appendix

A.1 Department of Education evidence tiers

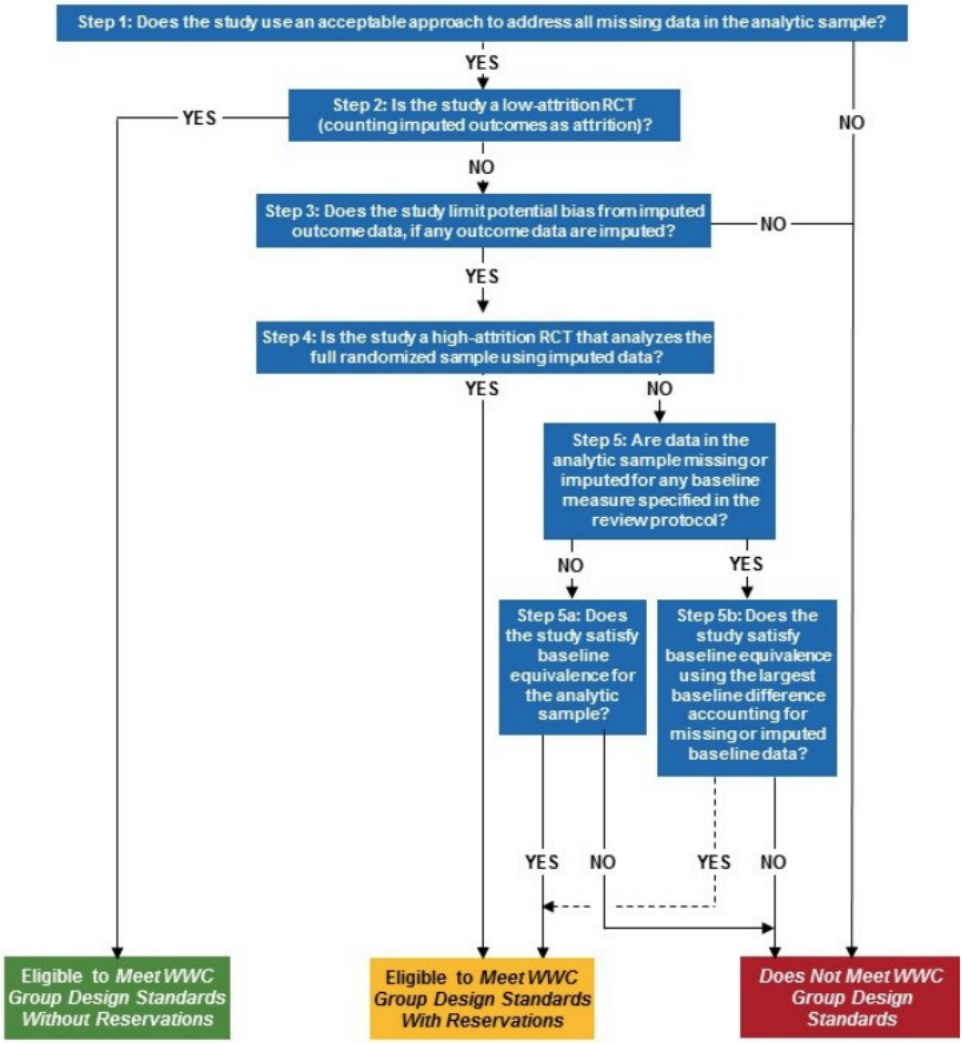
Figure II.1. Study Ratings for Individual-Level RCTs and QEDs



Note: To receive a rating of *Meets WWC Group Design Standards Without Reservations* or *Meets WWC Group Design Standards With Reservations*, the study must also satisfy the requirements in Chapter IV, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.

Figure A.1: DoE's tiers of evidence - implementation

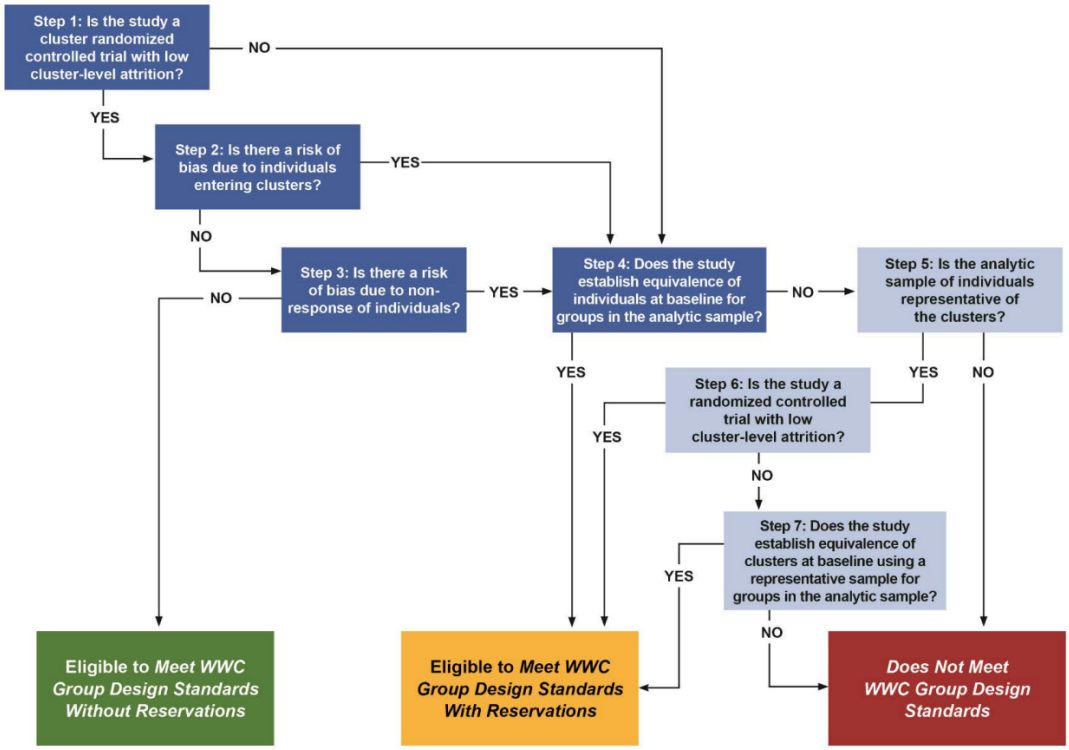
Figure II.5. Study Ratings for RCTs and QEDs with Missing Outcome or Baseline Data



Note: To receive a rating of *Meets WWC Group Design Standards Without Reservations* or *Meets WWC Group Design Standards With Reservations*, the study must also satisfy the requirements in Chapter IV, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.

Figure A.2: DoE's guidelines on missingness

Figure II.4. Review Process for Cluster-Level Assignment Studies



Note: To receive a rating of *Meets WWC Group Design Standards Without Reservations* or *Meets WWC Group Design Standards With Reservations*, the study must also satisfy the requirements in Chapter IV, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.

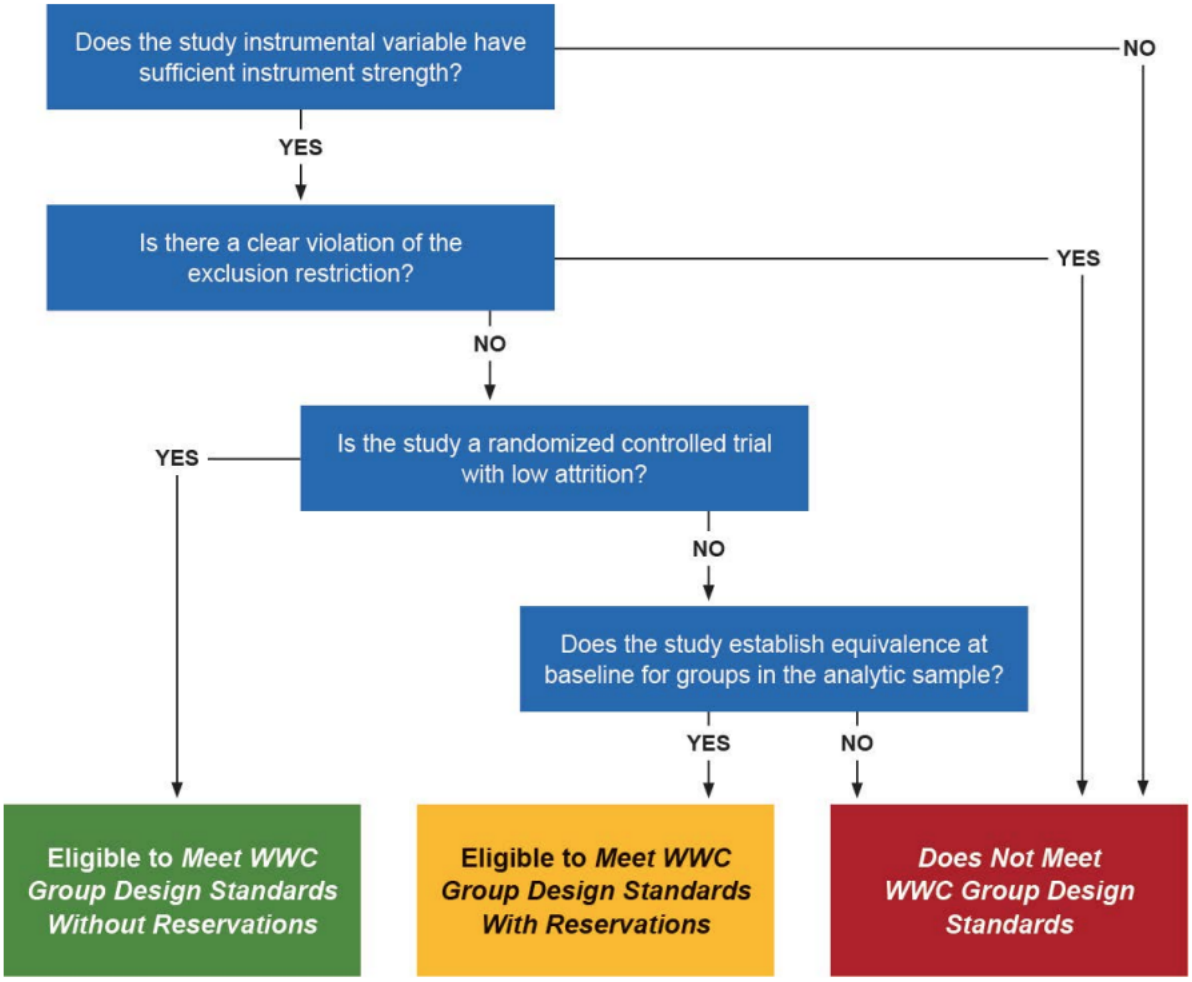
Figure A.3: DoE's guidelines on clusters

Table III.1. RDD Study Ratings

Standard	To be rated *Meets WWC RDD Standards Without Reservations*, studies must:	To be rated *Meets WWC RDD Standards With Reservations*, studies must:
1: Integrity of the forcing variable	Completely satisfy	Partially satisfy
2: Sample attrition	Completely satisfy	Partially satisfy at least one of these two standards
3: Continuity	Completely satisfy	
4. Bandwidth/Functional form	Completely satisfy	Partially satisfy
5. Fuzzy RDD	Completely satisfy	Partially satisfy

Figure A.4: DoE's guidelines on RDs

Figure II.6. Review Process for Studies that Report a CACE Estimate



Note: To receive a rating of *Meets WWC Group Design Standards Without Reservations* or *Meets WWC Group Design Standards With Reservations*, the study must also satisfy the requirements in Chapter IV, including that the study must examine at least one eligible outcome measure that meets review requirements and be free of confounding factors.

Figure A.5: DoE's guidelines on IVs and CACE

A.2 List of federal clearinghouses

Other federal agencies have followed in the DoE's footsteps and have started building databases on the existing research on specific policies. Not all of these databases have downloadable comprehensive databases, or evidence standards as clear and well-defined as the DoE's.

- Department of Labor (DoL)'s CLEAR's clearinghouse: evidence on on labor topics

- Corporation for National and Community Service (CNCS): evidence on what works in national service, social innovation, civic engagement, and volunteering
- U.S. Agency for International Development (USAID), YouthPower: evidence on what works in youth and peacebuilding, youth and health, youth and agriculture, food security, and nutrition
- US Departments of Agriculture and Defense's ClearingHouse for military family readiness: evidence on wide-ranging family and mental health issues.
- US Department of Health and Human services: multiple databases on programs whose purpose is to prevent and/or reduce delinquency or other problem behaviors in young people, teen pregnancy and substance prevention programs, etc.
- US Department of Justice: multiple databases on drugs and substance abuse, juveniles, crime and crime prevention, victims and victimization, law enforcement, technology and forensics, corrections and reentry, and courts