Evidence-Based Decisions and Education Policymakers

Nozomi Nakajima *

October 22, 2021

<Most recent version here>

Abstract

In today's era of evidence-based policymaking, education policymakers face pressure to use research to inform their decisions. This paper explores the mental models that policymakers use when integrating research evidence in their policy decisions, with a focus on education policymakers working in state and local education agencies in the United States. First, I examine policymakers' preferences for research evidence. Using a discrete choice experiment, I present policymakers with a series of research studies that vary along attributes of internal and external validity. I find that policymakers have preferences for larger studies and studies conducted in similar contexts as their own jurisdiction. However, they do not have a preference between experimental and observational studies. Second, I explore how much policymakers update their beliefs about the effectiveness of education policies using an information experiment. I show that policymakers update their beliefs in response to research evidence, but these effects are large and persistent only when the information presented contains a brief, accessible explanation of how the evidence was generated. The results of my study have implications for the production of education research, training of education leaders, and communication of scientific evidence.

*Ph.D. candidate, Harvard University (nnakajima@g.harvard.edu).

I thank my advisors David Deming, Felipe Barrera-Osorio, and Eric Taylor for support and advice. I am grateful to Peter Blair, Carrie Conaway, Emmerich Davies, Andrew Ho for helpful conversations and feedback, and to numerous seminar and conference participants for insightful discussions. I acknowledge generous funding and support from the James M. and Cathleen D. Stone fellowship from the Multidisciplinary Program in Inequality & Social Policy at Harvard University, the Research Officer's Grant from the William T. Grant Foundation, and a dissertation fellowship from the National Academy of Education/Spencer Foundation. This study was approved by Harvard's IRB (IRB19-1623) and pre-registered in the AEA (AEARCTR-0006563). All errors are my own.

1 Introduction

Education policymakers play an essential role in the design and implementation of education policies.[1] In today's era of evidence-based policymaking, policymakers face pressure to use research evidence to inform their decisions (Gordon & Conaway, 2020). This is particularly salient in the United States, where federal law mandates that education leaders implement policies, programs, and strategies that have been demonstrated to improve student outcomes (*Every Student Succeeds Act, Pub.L. 114–95*, 2015).[2] Despite the strong push to integrate research evidence in policymaking, we know little about the mental models used by education policymakers when making evidence-based decisions. In this paper, I examine policymakers' preferences for research evidence and how they update their beliefs in response to new information. I conduct survey experiments on over 2,000 policymakers working in state and local educational agencies in the U.S.[3] These policymakers are an important group to study because their decisions are consequential for how schools and teachers are organized and how students learn (Hightower, 2002; Spillane, 1996).

This paper addresses two research questions. First, **what preferences do policymakers have for research evidence?** Researchers studying education policies typically focus on establishing internal validity (Tipton & Olsen, 2018). However, the same policies can have different impacts for different populations (Deaton, 2010; M. Weiss et al., 2017) and policies that are effective in small trials may not be as effective when implemented at scale (Tipton, 2014; M. Weiss, Bloom, & Brock, 2014). These issues raise the question of external validity. Specifically, policymakers face the unique task of evaluating whether research findings are relevant to their specific local context. Prior studies show that education leaders often dismiss research because they "lack relevance" (Coburn & Talbert, 2006; Johnson et al., 2009) but we do not know how policymakers evaluate external validity—and specifically, relevance—of research evidence.

To study policymaker preferences for research evidence, I conduct a discrete choice experiment. Policymakers are presented with a series of research studies that vary along attributes of internal and external validity. They are asked about their preference between pairs of research studies as they make a hypothetical policy decision, requiring them to make tradeoffs between different study attributes. By estimating policymaker preferences for research evidence, I am able to reveal what types of evidence are likely to be used in policymaking.[4]

[1] I use the term "policy" to refer to policies, programs, and/or interventions, as the distinctions between them are not central to the motivation of the study.

[2] Given this policy landscape, I focus on instrumental use of research, where evidence directly shapes policy decisions. However, I acknowledge that there are other ways that research is used, such as symbolic/political use and conceptual use (C. Weiss, Bucuvalas, & Bucuvalas, 1980).

[3] I define policymakers as individuals in leadership roles, making policy/program decisions that affect multiple schools.

[4] Preferences elicited from choice experiments have been shown to closely correspond with real-world choices (Maestas, Mullen, Powell, Von Wachter, & Wenger, 2018; Mas & Pallais, 2017) and predict actual behavior (Hainmueller, Hangartner, & Yamamoto, 2015; Wiswall & Zafar, 2018).

In the discrete choice experiment, I find that policymakers have clear preferences for research evidence. They prefer studies with larger samples, multiple sites, and those conducted in similar settings as their own jurisdiction. This finding has important implications for the production of research. Across the social sciences, replication studies that re-evaluate, re-confirm, or extend findings of previous work are often not as highly regarded as novel studies with exciting results (Yong, 2012). In contrast to this current trend in academic publishing, my results show clear demand for research that tests the efficacy of policies in different contexts, as well as for replication studies that examine the effects of programs at a much larger scale.

At the same time, policymakers do not show preferences for the design of research studies. They place equivalent weight to experimental and correlational studies to inform their decisions. This result is surprising for several reasons. From a statistical perspective, experimental studies are less susceptible to threats of internal validity than observational studies. All else equal, experiments are the preferred research design if the goal is to advance policies shown to improve educational outcomes. From a policy perspective, federal guidelines under ESSA have established a "tier" of research evidence with experiments offering the strongest evidence for causal claims. Education policymakers are asked to follow these guidelines when deciding interventions to implement on the ground, yet my results suggest that they have different priorities in mind.

The second research question asks **how much do policymakers update their beliefs about the effectiveness of education policies?** Policymakers are implicitly asked to predict how well policies will work in their local contexts. But how do they form these predictions? Education research on policymakers often highlights the complexity of social and political processes in decision-making (Honig & Coburn, 2008), but few have empirically examined the cognitive aspects of belief formation (exceptions include Spillane (2000); Spillane, Reiser, and Reimer (2002)). This is a missed opportunity since policy implementation research beyond education points to ways in which cognition influences how policies get interpreted and enacted (C. Weiss et al., 1980).

To address this second research question, I elicit policymakers' predictions for the effect of an education policy in a particular setting. I then conduct an information experiment to study how they update their beliefs in response to new information from researchers and from education policymakers.[5] This experimental design is intended to mimic real-world scenarios; education policymakers have been known to seek out research evidence as well as views of their colleagues to inform their decisions (Penuel et al., 2017). Six weeks later, I follow-up with policymakers to examine if the information provision has persistent effects. The follow-up survey allows me to distinguish between true belief updating and experimenter demand or numerical anchoring. Using a Bayesian learning model, I analyze how policymakers respond to information

[5]This experimental design follows the recent economics literature on information-provision experiments in surveys (see Fuster, Perez-Truglia, Wiederholt, and Zafar (2018); Roth and Wohlfart (2020); Stantcheva (2020).)

signals from researchers and other policymakers to update their beliefs about the effectiveness of education policies.

In the information experiment, I find that policymakers are significantly more likely to update their beliefs when presented with information from researchers than from other policymakers. Moreover, they are significantly more likely to change their beliefs when research evidence is presented with accessible explanations of its research design. Even six weeks after the initial survey, the average policymaker places nearly 20% weight on the research evidence and 80% weight on their prior belief. This finding has important implications for scientific communication. Research findings are often communicated to policymakers as headlines of impact estimates with very little exposition about its research design (National Academies of Sciences, Medicine, et al., 2017; Schalet, Tropp, & Troy, 2020). My results suggest that policymakers are significantly more likely to incorporate research evidence in their decision-making process if they are able to follow and understand how these estimates were derived in the research presented to them.

My paper contributes to the literature on evidence-based policymaking in education, both in terms of generalizability and method.[6] The vast majority of studies on education policymakers have used case studies, providing rich detailed descriptions for a small number of units (Coburn, Honig, & Stein, 2009). Recent work by Penuel et al. (2017) is one of the few surveys on policymakers' use of research evidence, but their sample consists only of education leaders in the largest districts in the U.S. By partnering with an organization that offers professional development to education leaders, I am able to study 2,000 education policymakers who are tasked with making evidence-based decisions in districts and states across the country. Thus, findings from my study inform our understanding of the larger population of education policymakers in the U.S.

Methodologically, I use experiments to study how evidence-based decisions are made in education. This is important because decisions in education require a combination of values and evidence (Brighouse, Ladd, Loeb, & Swift, 2018). Policymakers must clarify what values are at stake to identify goals; then they must use evidence to evaluate whether a proposed policy will achieve those goals.[7] The link between values and evidence implies that observational data will naturally confound policymakers' values—which we can never fully observe—with the research evidence sought out and used by policymakers. However, by randomly assigning policymakers to see different research evidence (in the choice experiment) or to receive different types of information (in the information experiment), my experimental design ensures that the distribution or diversity of value judgments across policymakers is equal in expectation. This allows me to hold constant policymakers' values and hone in on how evidence affects their decision-making process.

[6] A handful of recent studies in economics examine aspects of evidence-based policymaking using experiments (see (Hjort, Moreira, Rao, & Santini, 2021; Rogger & Somani, 2018; Vivalt & Coville, 2020).

[7] Values encompass moral values like policymakers' judgment about the ideal goals of education, as well as social/political values like how to handle outside pressures that may limit the goals that can be achieved (Brighouse et al., 2018).

The remaining sections of the paper are organized as follows: Section 2 introduces the research design and surveys. Section 3 describes the sample of policymakers in my study. The conceptual framework, empirical strategy and results are described in Section 4 for the first research question and Section 5 for the second research question. The last section discusses the implications of the study and concludes.

2 Research Design

2.1 Recruitment

Participants were recruited from a non-profit organization that offers online professional development and training courses to educators working in state and local educational agencies in the United States. The organization enrolled participants through a rolling admissions process and offered asynchronous training courses. When participants enrolled in a course, they were invited to the main survey of the study and told that their responses will be used to inform the organization's future courses. Six weeks later, when participants completed their course, they were invited to a follow-up survey and told that their responses will be used to inform policy priorities in their own education agencies. As a result, my study setting encourages participants to (i) complete the surveys because they are invited during the course enrollment and completion processes, and (ii) reveal their true preferences because their responses may inform decisions made by the professional development organization and their own education agency. Participants were informed that participation in the study was voluntary and that they would enter a raffle to receive a discount voucher for future professional development courses.

The main survey was deployed between October 2020 and April 2021, and the follow-up survey was administered between November 2020 and June 2021. Of the 2,567 education policymakers invited to the study, 2,245 completed the main survey (87%), and 2,079 completed both the main survey and follow-up survey (81%). Item non-response did not exceed 1% for any question in the survey. Attention check questions were passed by all but six respondents. The median total survey time was 35 minutes for the main survey and 7 minutes for the follow-up survey.

2.2 Survey Structure

In this section, I describe the overall structure of the main survey and follow-up survey. Additional design details of the survey, including screenshots of the instructions and tasks, are included in the Online Appendix.

2.2.1 Background

The main survey begins with a background questionnaire about the policymaker. To capture the context of where they work, I ask respondents to state their best guess for (i) the percent of students receiving free or reduced-price lunch in their jurisdiction and (ii) the percent of white (non-Hispanic) students in their jurisdiction. The responses to these two questions can be compared against the Common Core of Data from the National Center for Education Statistics, which allow me to measure the accuracy of their responses. By benchmarking the policymakers' responses to actual data, I have a proxy measure for how well they know their own context.

Next, I measure how well policymakers can evaluate scientific research using the Scientific Reasoning Scale (SRS). The items of the SRS ask respondents to apply their reasoning skills to brief scientific scenarios. Previous studies have shown that individuals who score higher on the SRS exhibit higher levels of numeracy and cognitive reflection, and perform better on tasks requiring analysis of scientific information (Drummond & Fischhoff, 2017). Together, the context questions and SRS provide baseline measures of policymakers' ability to evaluate external validity and internal validity of research.

2.2.2 Discrete choice experiment

The second section in the main survey contains a discrete choice experiment. Respondents are asked to evaluate different research evidence to help guide policy decisions in their own jurisdiction. Each task contains two potential research studies, which randomly vary along aspects of internal and external validity with the intent of creating realistic variation of study attributes. Specifically, each study varies along six attributes that are most commonly reported in evaluations of education policies and programs (Orr et al., 2019): (1) research design, (2) sample size, (3) number of sites, (4) percent of students receiving free or reduce price lunch, (5) percent of non-Hispanic White students, and (6) urbanicity. The table below shows the possible levels of each of the six attributes, which were based on actual research studies to keep the task realistic for policymakers (see Online Appendix C).

In the discrete choice experiment, the levels of attributes vary randomly, with randomization occurring independently across respondents, tasks, and attributes. In order to avoid confusion, the order in which attributes appear in tables is fixed across tasks for each individual respondent. Each respondent completes the task five times, evaluating a total of 10 potential research studies.

Below each table, respondents are asked two questions to measure their preference. They are asked to: (i) make a forced choice between the two studies (*"If you had to choose one of the two studies, which study is more useful...?"*) and (ii) consider both studies and rate each using a constant sum scale of 100

Attribute	Levels
Research design	Description of observational study Description of experimental (lottery) study
Sample	500 students 2000 students 15000 students
Sites	1 site 10 sites 25 sites
Poverty (% students receiving free- or reduced-price lunch)	+/- 5 percentage point from own district +/- 25 percentage point from own district +/- 45 percentage point from own district
Race (% white non-Hispanic students)	+/- 5 percentage point from own district +/- 25 percentage point from own district +/- 45 percentage point from own district
Urbanicity	Urban Suburban Rural Mix of urban, suburban, and rural

(*"If you had to consider both studies, what weight would you assign to each study...?"*). The latter question gives respondents greater flexibility by allowing them to incorporate multiple pieces of research evidence for decision-making and to accept the possibility of no preference (i.e., 50:50 weighting) between two studies. The order of these two questions is randomized at the respondent level.

In the discrete choice experiment, policymakers are asked to evaluate research on charter schools. This topic is chosen for two key reasons. First, there is a large repository of charter school effectiveness studies with considerable variation along the six attributes studied in this paper. For example, charter schools have been evaluated using both lotteries (Angrist, Pathak, & Walters, 2013; Chabrier, Cohodes, & Oreopoulos, 2016; Furgeson et al., 2012; Gleason, Clark, Tuttle, & Dwoyer, 2010) and observational studies (Center for Research on Education Outcomes, 2009, 2015; Zimmer, Gill, Booker, Lavertu, & Witte, 2012). They have also been evaluated in urban settings like Boston (Cohodes et al., 2013) and in rural areas (Furgeson et al., 2012). This variation in the design and context of existing research on charter school is important because it makes the discrete choice experiment a realistic exercise for policymakers.

Second, charter schools are a contested, polarizing policy issue among educators that defies partisan division (Cheng, Henderson, Peterson, & West, 2019; Kirst, 2007). As a result, charter school policy presents an important opportunity to understand preferences for scientific evidence when policymakers are likely to hold strong views about an issue. Importantly, the findings from the research studies are not presented in the choice experiment. This ensures that preferences are not driven by confirmation bias or motivated reasoning, but rather by attributes related to the internal and external validity of research.

2.2.3 Information experiment

The information experiment has four parts. First, I measure policymakers' prior beliefs by asking them to predict the effect of an education policy. Then, policymakers are asked to rank their choices between different pieces of information that may help them form more accurate beliefs. The choices are: effect size predictions made by peer policymakers, effect size estimates by researchers, or no information. In the third part of the experiment, policymakers are randomly assigned to one of these pieces of information and their posterior beliefs are elicited. Six weeks later, policymakers' posterior beliefs are measured again in a follow-up survey. This final part is used to examine if the information provision has any persistent effects on belief updating.

Prior beliefs. First, I measure policymakers' prior beliefs. Policymakers are asked to guess the effect of an education policy in a particular setting: the expansion of charter schools in an urban school district. The prompt in this belief elicitation task describes the Boston charter school expansion evaluated in Cohodes, Setren, and Walters (2021):

> *An urban school district in the United States is seeking guidance from education leaders like you. The district is considering whether to expand its charter school sector. They want you to help predict the effect of charter schools on the math test scores of students in their district who attend charter schools. 12% of students in the district are White (non-Hispanic) and 84% of students in the district receive free or reduced price lunch. The charter schools in this district have high expectations for their students. Traditional public schools in this school district have relatively low math test scores.*

The task is designed to capture the implicit decision-making task of policymakers, which is to predict the effect of education policies or programs in specific contexts. It is worth noting that the setting described in this task is not personalized to each policymaker's local jurisdiction because of data limitations. In the analysis, which will be described in more detail in Section 5, prior beliefs are benchmarked to real, estimated effects of charter schools. Unfortunately, it is impossible to know the effect of charter schools (or any education policy/program) in every jurisdiction represented by the policymakers in my sample. As a result, the same setting is described to all policymakers. Despite this design limitation, the task is still relevant because education leaders frequently change jobs and locations (Grissom & Andersen, 2012), which suggests that they can be required to make decisions in relatively new and unfamiliar settings.

To elicit subjective probability distributions for the effect of urban charter schools on student achievement, I ask respondents to freely select three support points (in this case, effect sizes) and then assign probabilities to each. This approach to eliciting subjective beliefs, which is modeled after Altig et al. (2020), has two

key advantages. First, it gives flexibility to the respondent. Respondents are able to express the range of expected effects, the uncertainty of these expected effects, and any skewness in the distribution of these effects. Second, it avoids anchoring effects associated with pre-specified support points.[8] As a result, I am able to capture the heterogeneity across policymakers' prior beliefs about the effect of urban charter schools.[9]

In this elicitation task, it is important that policymakers understand and can express their beliefs in terms of effect sizes and probabilities. To ensure that policymakers in my sample can do so, they work through two "warm up" exercises in the beginning of this section of the survey. First, respondents answer an opening question to familiarize them with answering probabilistic questions. Second, respondents see an interactive data visualization that familiarizes them with interpreting effect sizes and understanding the range of effect sizes that have been reported in education research.[10]

Information selection. After the elicitation task, respondents are asked to rank their choices between different pieces of information that are potentially relevant to improving their prediction. The options are (i) the effect of urban charter schools predicted by peer policymakers, (ii) the effect of urban charter schools estimated by researchers, and (iii) no information. The design of this question is motivated by prior research showing that education leaders often seek out information from both educators and researchers when making decisions (Penuel et al., 2017). Specifically, the purpose of this question is to understand policymakers' information preference when they are trying to form accurate policy predictions.

Information treatments. Later in the survey, respondents are randomly assigned to one of four conditions. The first condition receives no information (control). The second condition (policymaker) sees the forecast made by other policymakers who completed the same prediction task in a pilot survey conducted in February-March 2020. Respondents who are randomly assigned to the peer policymaker condition are told that other policymakers predict an effect of 0.04 s.d. and that they are 95% confident that the effect is between -0.04 s.d. and 0.12 s.d. The third condition (researcher) sees the forecast made by researchers, which is the average treatment effect estimate from experimental studies of urban charter schools reported in (Chabrier et al., 2016).[11] Policymakers randomly assigned to this condition are told that researchers predict an effect of 0.25 s.d. and that they are 95% confident that the effect is between 0.16 s.d. and 0.34 s.d. Finally, the

[8]The survey items were piloted in February-March 2020 and a subset of participants from the pilot were part of a focus group to assess comprehension of the questions. Instead of the five-point subjective probability distribution in Altig et al. (2020), I use a three-point subjective probability distribution in the interest of time while acknowledging that it offers a coarser characterization of the subjective probability distributions.

[9]It is unlikely that respondents cheated in this elicitation task. First, the prompt did not mention the Cohodes et al. (2021) study, so it is unlikely that they would even know to look up the paper that inspired this prompt. Second, respondents were asked at the end of the survey if they had looked up any outside sources when completing the survey. All respondents answered no.

[10]See Figure B.4 for details of these warm-up exercises.

[11]Importantly, the estimates reported in Chabrier et al. (2016) was published before the charter school expansion study in Cohodes et al. (2021). This allows the researcher information treatment to capture what researchers would have forecasted in the prediction exercise.

fourth condition (researcher-plus) sees the same estimate as the researcher treatment group but also includes an accessible explanation about the research design to help respondents understand how the researchers derived these estimates. After random assignment to one of the four conditions, respondents are given the opportunity to update their predictions based on the information shown to them. To measure the posterior beliefs of policymakers, the exact same task to elicit subjective probability distributions is presented again.

Follow-up survey. Six weeks after the survey, respondents receive a follow-up survey. This follow-up survey addresses four key concerns with the information experiment. First, it aims to mitigate experimenter demand effects. By presenting the follow-up survey as an independent study from the initial survey, it is unlikely that policymakers assigned to different information treatment arms in the initial survey will make different inferences about the experimenter's expectations in the follow-up survey. I obfuscate the connection between the main survey and the follow-up survey by using different layouts for the survey invitation and consent forms, and include a question unrelated to the information experiment in the beginning of the follow-up survey (Haaland, Roth, & Wohlfart, 2020).

Second, the follow-up survey aims to address concerns that survey experiments do not capture actual behavior. In this survey, respondents are given an opportunity to provide policy recommendations to their local education agency and they are told that their recommendations will be considered for strategic decisions by the agency. This policy recommendation process is similar to Liaqat (2019) and designed to increase the real-world stakes associated with taking the survey. In the policy recommendation process, respondents are asked to rank different education policy issues that are most pressing to their setting.[12]

Third, the follow-up survey includes an open-ended question to alleviate concerns that policymakers may be primed by available answer categories. Respondents are asked, "What has informed your policy views about the effectiveness of charters at improving student achievement?" This open-ended question allows me to understand how policymakers' rationalize their views about charter school. The key advantage with this format is that respondents are not primed by available answer choices, allowing me to directly measure the first idea that comes to mind (Bursztyn, Haaland, Rao, & Roth, 2020; Stantcheva, 2020).

Finally, this follow-up survey is designed to address concerns about numerical anchoring from the initial survey. In the very last question of the follow-up survey, I re-elicit policymakers' posterior beliefs. This elicitation task is exactly the same as the initial survey, which allows me to examine whether any effects

[12]This policy recommendation process is real. There are several local education agencies (LEA) that regularly send their employees to attend the partnering organization's professional development courses. These LEAs routinely ask the partnering organization for aggregated data to understand how their participants benefit from the courses. These LEAs agreed to also consider policy recommendations from their participants. This agreement was established separately between the partnering organization and specific LEAs so as the researcher, I do not know which local education agencies actually agreed to and therefore received the policy recommendations. To be honest to survey respondents, the questionnaire says that several local education agencies are participating in the policy recommendation process and that a respondent's local education agency may be participating (see Figure B.7).

of the information experiment persists six weeks later. At this point, respondents are likely to make a connection between this follow-up survey and the initial survey but this question is presented as the last item in the follow-up survey.

3 Sample characteristics

3.1 Demographic information

To set the stage for the rest of the paper, this section provides information about the sample of policymakers in this study. Background information on each respondent comes from enrollment records collected by the partnering organization. These administrative records contain information about the policymaker's job (name and ZIP code of the policymaker's current and previous employer, current job title, and primary job tasks) and demographic information (gender and race). The information in these records were verified by each respondent's employer during enrollment, ensuring that policymakers in my sample can be identified to the correct jurisdiction.

Table 1 Panel A shows that my sample is composed of four types of policymakers. District leaders (43.2%) and state leaders (6.6%) are individuals in top leadership positions at the district or state level, holding titles such as superintendent, chief schools officer, and chief executive officer. District administrators (35.4%) and state administrators (14.7%) are responsible for departments within district and state education offices, such as chief development officers, chief information officers, and curriculum directors. In my sample, 10.7% of education policymakers stated that their primary job task involved data or research related tasks. These policymakers held job titles such as director of research and evaluation, director of data and assessment, and chief of data systems and research.

Panel B presents summary statistics for my sample compared to a nationally representative sample of K-12 public school teachers (American Teacher Panel, ATP) and school leaders (American School Leader Panel, ASLP). Although these nationally representative samples do not include district and state-level education policymakers, they describe the pipeline of education leaders in the country and provide a benchmark for understanding my sample. My sample is mostly female (79.5%), which is similar to the teacher workforce but has considerably more female representation than the school leadership sample. The policymakers in my sample are more racially diverse (62.9% white, 14.0% black, 10.3% Hispanic, and 6.7% Asian) than the teacher or school leader samples. In terms of where they work, the majority of policymakers in my sample work in urban settings (56.4%) with an average student body population comprised of 41.5% white students and 51.8% eligible for free- or reduced-price lunch. Figure A1 shows that the policymakers were

distributed across 49 states, Washington, D.C. and Puerto Rico. Overall, policymakers in my sample are more racially diverse and work in settings with more representation of racial minorities than the average American education workforce.

3.2 Context accuracy and scientific reasoning

Figure 1 summarizes how well policymakers know their own policy context. Panel A visualizes the percent of students eligible for free- or reduced-price lunch in a policymaker's jurisdiction and Panel B visualizes the percent of white students in a policymaker's jurisdiction. The horizontal axis is the actual data reported by the National Center for Education Statistics (NCES) and the vertical axis is the policymaker's guess of the data. If policymakers have perfect accuracy, we would expect to see a straight, 45-degree line of the scatter plots, as there would be no deviation between the two measures. Overall, Figure 1 shows that policymakers have quite accurate estimates of their own context as measured by median absolute deviations. The median policymaker's estimate deviates from the NCES data by 7.50 percentage points for the percent of students eligible for free- or reduced-price lunch and by 6.54 percentage points for the percent of white students.

Next, Table 2 summarizes how well policymakers can reason through scientific concepts. On average, policymakers correctly answered between 7 and 8 out of the 11 items on the Scientific Reasoning Scale (SRS). Given the importance of causal inference for evaluating programs and policies, it is worth noting that many – but far from all – policymakers understand causality (63.8%), confounding variables (69.8%), control group (80.4%) and random assignment (75.5%). To benchmark these numbers, columns 2 and 3 present the summary statistics from the validation study of the SRS (Drummond & Fischhoff, 2017). The samples from the validation study are MTurk participants, which are mostly young, white, college-educated Americans. Policymakers in my study tend to score higher overall, but appear to have weaker understanding of confounding variables than these MTurk samples. As another point of reference, (Hill & Briggs, 2020) recently surveyed principals and central office staff from mid- and large-sized U.S. urban districts. The authors develop survey items that measure educators' knowledge of ideas from statistics and research design. In their sample, the authors find that few educators understand internal validity (33%) and random assignment (46%). While items from the SRS cannot be directly compared to items from Hill and Briggs (2020), it is worth noting that the majority of policymakers in my sample demonstrate some understanding of concepts related to causal inference.

4 Policymaker preferences for research evidence

4.1 Conceptual framework and empirical strategy

The first research question in this paper asks: what research evidence do policymakers prefer? Answering this question presents a couple of empirical challenges. First, there are limited data on what research evidence is selected and used by education policymakers. Second, a key shortcoming of existing data is that characteristics of research evidence are likely to be confounded with other factors, making it difficult to isolate and interpret policymaker's preferences. For instance, schools included in randomized controlled trials are not representative of public school in the U.S. (Stuart, Bell, Ebnesajjad, Olsen, & Orr, 2017; Tipton et al., 2016). If existing data on policymakers' use of research evidence contains information about the research design (i.e., whether a study uses a randomized experiment), but not about other school features that are correlated with policymaker preferences, we would expect the estimates of research preferences to be biased.

I overcome this challenge by experimentally manipulating the hypothetical studies provided to policymakers in the discrete choice experiment described above. By randomizing the levels of study attributes, the survey decreases bias from the correlation of observed and unobserved study characteristics. However, the six attributes presented are not exhaustive of all possible study characteristics, and it is possible that policymakers may still make inferences about unobservable attributes. To minimize this possibility, the survey explicitly instructs that the studies presented differ only in the attributes provided and are otherwise identical.[13]

In the discrete choice experiment, each policymaker completes the choice task five times. Each choice task consists of a hypothetical policy scenario and two studies. The policymaker selects between two possible studies to inform their policy decision. Thus, the resulting dataset contains $5 \times 2 = 10$ unique observations for each policymaker. To estimate preferences, I regress policymaker i's selection of study j (for the forced choice response) or percent weight assigned to study j (for the rating response) on a vector of study attributes shown to respondents in the choice task $X_j = [X_{j1}, ..., X_{j6}]$ as follows:

$$u_{ij} = X_j'\beta + \varepsilon_{ij} \tag{1}$$

The coefficient of interest β is the average marginal conditional effect; it is the average effect of a study attribute on policymakers' preferences when they are also given information on the other five study attributes. The standard errors are clustered at the individual policymaker level. For ease of interpretation, equation

[13] It is still possible that policymakers may not internalize this instruction. For example, they may believe that an attribute presented in this survey (like sample size) is correlated with other aspects of study attributes that are not included (like academic discipline of the researchers conducting the study). As noted in Wiswall and Zafar (2018), biases like these make discrete choice experiments similar to audit studies in their limitations.

(1) is estimated using a linear probability model for the forced choice response. As a robustness check, I also estimate the equation using logistic regression. For the rating response, equation (1) is estimated using linear regression. In Table A.1, I show that the main results are robust to these different model specifications.

The key assumption for identifying preferences is that unobserved study characteristics are independent of the experimentally manipulated study attributes. To assuage concerns about the implementation of the randomization process in the discrete choice experiment, I perform several diagnostics. I check that policymaker preferences are not driven by how the studies are presented. Specifically, I re-estimate equation (1) with the following interaction terms: attributes $\times$ task number (1-5), attributes $\times$ study order (A/B), and attributes $\times$ order of attributes (1-6). Then, I perform an F-test for the joint significance of the interaction terms. Because the discrete choice experiment randomizes the levels presented in tables independently across respondents, across tasks, and across attributes, I expect all of these interaction terms to yield null effects.

The results of these robustness checks are presented in Table A.2 columns (1) through (3). None of these interaction terms are statistically significant, which means that policymaker preferences captured in the discrete choice experiment are not influenced by the task number, the order of the study, or the order of the attributes presented.

I also check that policymaker preferences are not strictly determined by how they are measured. By design, I elicit their preferences in two ways (forced choice and rating) and randomize the order that the questions are presented to each respondent. Therefore, I can re-estimate equation (1) with attributes $\times$ question order to confirm that the order of preference elicitation is not affecting my results. The results of this robustness check are presented in Table A.4 column 4. None of the interaction terms are statistically significant, confirming that the order in which the two questions were asked did not affect policymakers' preferences.

Another measurement concern is that idiosyncratic features of the studies—instead of the experimentally manipulated attributes—are affecting my results. To check that this is not the case, I include a placebo in the choice task. As shown in Appendix A, the columns of each table are randomly shaded either in light-blue or bluish-gray. This placebo attribute should have no impact on preferences if policymakers are following the survey instructions and making their decisions based on only the presented attributes. In Table A.2 column 5, I show that there is no effect of the placebo on policymaker preferences.

4.2 Results

Figure 2 summarizes the estimated average marginal conditional effect for each of the attributes included in the discrete choice experiment. The left figure shows the results for the forced choice outcome and the right

figure presents the results for the rating outcome. The key results are broadly the same regardless of how the outcome is measured.

Several findings are notable from Figure 2. First, presenting experimental studies rather than correlational research has a small and statistically insignificant effect on policymaker preference for research evidence. From a researcher's perspective, this is a surprising finding given the importance of research designs when advancing causal claims about programs and policies. All else equal, experimental studies are less susceptible to threats of internal validity than observational studies, making experiments the preferred research design. The lack of policymaker preference for research design is also surprising in light of the federal policy surrounding the use of research evidence. The Every Students Succeeds Act (ESSA), which specifies tiers of evidence that policymakers should consider when making recommendations, makes clear that experiments offer the strongest evidence for causal claims. Despite this explicit guidance, policymakers do not seem to show preference for research designs when using research to inform their policy decisions.

Why don't policymakers prefer experimental evidence? The results may be partly explained by policymakers not fully internalizing the value of experiments. While the guidance under ESSA familiarizes policymakers with keywords like "randomized control trials" and "experiments", policymakers may not necessarily comprehend the value of these methods. In the discrete choice experiment, the research design attribute has two levels: a description of an observational study (*compared students attending charter schools with students attending traditional public schools*) and a description of a lottery study (*compared students who were offered a seat to charter schools with those not offered a seat to charter schools, based on a lottery*). This design requires policymakers to understand that randomly assigning students to schools through a lottery overcomes the selection bias that threatens observational studies. Consistent with this explanation, I find that policymakers who have higher scientific reasoning skills are significantly more likely to prefer the experimental study. Table 3 column 1 shows that the probability of choosing an experimental study increases by 1.3 percentage points ($p < 0.01$) for every item correctly answered on the Scientific Reasoning Scale. I also find that policymakers whose primary job function involves data and research-related tasks prefer experimental studies. These policymakers are 4 percentage points more likely to select experimental studies than other policymakers. However, we cannot reject the null hypothesis that their preference for research design is the same as other policymakers given the large standard errors arising from smaller samples of policymakers in these data and research roles.

The second key result is that policymakers have strong preferences for bigger studies – both in terms of sample size and number of sites. Changing the sample size from 500 to 2,000 increases the probability that policymakers consider a research study to inform their policy decision by 10.9 percentage points. Similarly, policymakers are 14.0 percentage points more likely to prefer a research study that includes 10 sites rather

a study conducted in one site. While larger studies are generally preferred, policymaker preferences do not scale linearly. Figure 2 makes clear that increasing the sample size from 500 to 2,000 students has a larger effect on preferences in terms of effect-per-student than changing the sample size from 500 to 15,000 students. The same is true for increasing the number of sites. This non-linearity is good news for researchers; studies do not need to be onerously large-scale for policymakers to pay attention.

Policymakers' preference for larger samples and more sites is consistent with the methodological view of generalizing "from broad to narrow" (Cook, Campbell, & Shadish, 2002). In order to assess whether findings from a study will hold out-of-sample, researchers often frame external validity in two ways. The first uses study estimates to make inferences about impacts in a larger population from which the study sample is selected (narrow to broad generalization). The second way – which policymakers in my sample follow – involves estimates from multiple sites to predict the impact in another site that is not part of the study sample (broad to narrow generalization).

The third main finding is that policymakers prefer studies conducted in settings with similar poverty rates, urbanicity, and racial composition as their own local jurisdiction. Policymakers are particularly sensitive to the percent of students eligible for free- or reduced-price lunch (FRPL), a proxy measure for poverty that is widely used in education research. The coefficients for this attribute appear to grow linearly; for every 1-percentage point difference between a study's and a local jurisdiction's percent of FRPL eligible students, the probability that a study is selected decreases by 0.3 percentage points. While policymakers are highly sensitive to small changes in poverty rates, they care about changes in racial composition in broader terms when deciding which research to inform their policy decisions. Policymakers are significantly less likely to rely on studies with very dissimilar racial composition (+/- 45 percentage points) as their local context, but they seem to be willing to consider evidence conducted in settings with somewhat different racial composition (+/- 25 percentage points). Finally, urbanicity is also an important attribute for policymaker preference. Studies conducted in congruent contexts as a policymaker's local jurisdiction is 7.9 percentage points more likely to be considered for informing their policy decision.

5 How do policymakers update beliefs about the effectiveness of education policies?

This section presents results from the information experiment, which examines how policymakers update their beliefs about the effectiveness of education policies. My analysis proceeds in four parts, following the stages of the experiment outlined in Section III. First, I present policymakers' prior beliefs about the effect

of urban charter schools. Second, I examine what information sources rank highest among policymakers, and whether there is heterogeneity in their preferences. Third, I study how policymakers use (and don't use) information from peer policymakers and researchers by leveraging the random assignment of different information sources in the survey. Finally, I examine if the information provision has persistent effects by following-up with policymakers six weeks after the initial survey.

5.1 Prior beliefs

To measure prior beliefs about the effect of urban charter schools on student achievement, respondents selected support points and then assigned probabilities to each. Each policymaker reported three support points $\{Effect\}_{i=1}^{3}$, with associated probabilities $\{p_i\}_{i=1}^{3}$. Then, I calculate policymakers' subjective mean and variance as:

$$Mean(Effect) = \sum_{i=1}^{3} p_i \times Effect_i \tag{2}$$

$$Var(Effect) = \sum_{i=1}^{3} p_i \big(Effect_i - Mean(Effect)\big)^2 \tag{3}$$

Figure 3 shows the histograms of the mean and variance of policymakers' prior beliefs. On average, policymakers believe that urban charter schools described in the prompt only do marginally better than traditional public schools, with a mean effect size of 0.0312 standard deviations (s.d.). Policymakers express considerable uncertainty in their prior beliefs, with a mean variance of 0.856 s.d.

Of the 2,079 policymakers in my sample, 77 reported mean prior beliefs below the 2nd percentile (-1.3 s.d.) or above the 98th percentile (1.4 s.d.). In my analyses, I drop these respondents as these extreme beliefs may reflect typos or lack of attention. Trimming my sample should not affect the experimental analysis since the information treatment occurs after eliciting policymakers' prior beliefs. However, policymakers may also express extreme values in their post-treatment outcomes. Instead of trimming the sample based on post-treatment outcomes, I winsorize their posterior and follow-up beliefs to a minimum of -1.3 s.d. and maximum of 1.4 s.d.

5.2 Information selection

Policymakers are asked what information sources would be most useful to accurately predict the effect of urban charter schools. They are presented with a choice between what researchers predict, what peer policymakers predict, and no information. The bottom row in Table 4 summarizes the most preferred

source of information. 59.7% of policymakers chose researcher forecasts, 34.0% preferred peer policymaker forecasts, and 6.3% chose no information. These results suggest that policymakers have considerable interest in knowing what researchers think when asked to make accurate policy predictions.

Importantly, these information sources vary in terms of how useful or informative they are for making accurate predictions. I define usefulness as having low absolute difference in means between the information signal and the actual estimate of the policy. As mentioned in Section III, the actual estimate of the policy reported in (Cohodes et al., 2021) is 0.22 s.d., whereas the estimate by peer policymakers is 0.04 s.d. and the estimate by researchers is 0.25 s.d. Based on the absolute difference in means, the researcher forecast is more useful (with an absolute difference of 0.03 s.d.) than the policymaker forecast (with an absolute difference of 0.18 s.d.). Thus, the preference for information by the average policymaker is consistent with the usefulness of the information signals.[14]

In Table 4, I examine the heterogeneity of information preferences. Each cell represents the bivariate relationship between the information most preferred (column) and a covariate (row). Two key findings emerge from this analysis. First, policymakers with higher scientific reasoning skills are systematically more likely to rank researcher forecasts on top, followed by policymaker forecasts, then no information. This suggests that there are differences in demand for information along scientific reasoning skills. Second, policymakers with larger variance in their prior beliefs are most likely to prefer research forecasts, followed by policymaker forecasts, then no information. Thus, policymakers who are less confident in their predictions about the effectiveness of urban charter schools exhibit stronger demand for information.

5.3 Information treatments

5.3.1 Conceptual framework and empirical strategy

To study how policymakers update their beliefs about policy effectiveness, I use a Bayesian learning model. Policymaker i has a prior belief about the policy's effect $\theta_i^{prior} \sim N(\mu_i^{prior}, v_i^{prior})$, where μ_i^{prior} is the mean of i's prior and v_i^{prior} is the variance of their prior. The policymaker receives information related to the policy $\theta_i^{inf} \sim N(\mu^{inf}, v_i^{inf})$, drawn from a distribution centered around μ^{inf} with a perceived uncertainty of v_i^{inf}. Then, we expect the policymaker's posterior belief θ_i^{post} to be a weighted average of the information acquired and their prior belief, with weights proportional to the relative precision of each component:

$$\theta_i^{post} = (1 - \pi)\theta_i^{prior} + \pi\theta_i^{inf} \tag{4}$$

[14]This pattern may not generalize beyond this task, as it is possible that policymakers may rank the forecast from individual policymakers who they know and trust to be higher than the forecast from an average researcher.

where $\pi = (v_i^{prior})/(v_i^{prior} + v_i^{inf})$. That is, π can take a value from 0 (policymaker ignores the information) to 1 (policymaker fully adjusts to the information). The equation can be re-arranged so that:

$$\theta_i^{post} - \theta_i^{prior} = \pi(\theta_i^{inf} - \theta_i^{prior}) \tag{5}$$

This shows that the slope between $(\theta_i^{inf} - \theta_i^{prior})$ and $(\theta_i^{post} - \theta_i^{prior})$ can be used to estimate the rate at which policymakers update their beliefs. However, policymakers' posterior beliefs could revise toward the information even if they have not seen it, so I need to separate true learning from spurious reversion to the information. To do so, I leverage the random assignment to different information sources in my experiment and fit the following regression specification:

$$\theta_i^{post} - \theta_i^{prior} = \pi(\theta_i^{inf} - \theta_i^{prior})I_i + \gamma(\theta_i^{inf} - \theta_i^{prior}) + \rho I_i + \varepsilon_i \tag{6}$$

where I_i is a categorical variable indicating the type of information received in the experiment. The coefficients of interest are π, which capture the true learning rate (relative to the control group) and γ is the degree of spurious mean-reversion. To examine whether belief updating varies between different information sources, I compare the coefficients of π for different values of I_i.

I also re-estimate equation 6 using θ_i^{post} measured at the follow-up survey. If updating behavior from the information experiment is driven only by experimenter demand or numerical anchoring and not by real learning, we should not expect persistent effects six weeks after the information was delivered.

5.3.2 Balance and attrition

Before turning to the main results of the information experiment, I alleviate concerns related to the experimental design. To check that the information was actually randomly assigned, I begin by examining the baseline covariates across the three treatment arms and control group. Table A.3 shows balance in policymaker demographics as well as in the settings where policymakers work. Only one out of the 48 tests is statistically significant at the 10% level or less, consistent with what we would expect by random chance. Figure A.2 also shows the prior beliefs of policymakers across the four information groups. As expected from randomization, the distributions are not significantly different from one another.

A key threat to identification of the information treatment effect is differential attrition between different information conditions. To address this concern, Table A.4 presents the attrition rates of the control group, policymaker group, researcher group, and researcher-plus group. The rate of non-response ranges between 6.5% and 8.4%, and these differences are not significantly different from one another. I also examine if the mean of the baseline covariates differ across groups for respondents who started the survey. Overall, I do

not find evidence of selective attrition based on observable characteristics.

5.3.3 Main results

Figure 4 summarizes the key results of the information experiment. The y-axis is $\theta_i^{post} - \theta_i^{prior}$ in equation (6), which captures the revision in policymaker beliefs. The x-axis is $\theta_i^{inf} - \theta_i^{prior}$ in equation (6), which is the gap between the information signal and prior beliefs. For ease of interpretation, I plot the regression line (with 95% confidence interval) for each group separately. If policymakers fully updated their beliefs to the information shown, we would expect a slope of 1 (45 degree line).

The first key finding from Figure 4 is that policymakers are significantly more likely to update their beliefs about the effectiveness of urban charter schools when presented with information from researchers than from other education policymakers. Panel (a) shows a slope of 0.248 for those who saw what other policymakers predicted. This means that the average policymaker placed 24.8% weight on the information from other policymakers and 75.2% weight on their prior belief. In contrast, the average policymaker placed 35.5% weight on the researcher estimate. The effect is similar in the researcher-plus group, with a weight of 34.3% on the information received. This finding is important, as it suggests that policymakers place considerable weight on research evidence to form their beliefs about the effectiveness of education policies.

The second key finding is that policymakers are significantly more likely to change their beliefs when research evidence is presented with accessible explanations of the research design. As expected from information dilution over a six-week period, the effect of information declines in the follow-up survey. Panel (b) shows that the average policymaker only placed 7.2% weight on researcher estimates in the follow-up survey, which suggests that the initial reaction to the information was mostly spurious. In contrast, the effect of receiving the researcher-plus treatment decreases over time but remains substantial, with a weight of 20.2%. Together, these results suggest that providing policymakers with research evidence can change their beliefs about the effectiveness of education policies. Moreover, policymakers are more likely to change their beliefs when research evidence is presented with accessible explanations of the research design underlying the evidence.

5.3.4 Heterogeneity of learning rates

Figure 5 presents examines heterogeneity in learning rates between different policymakers. In the Bayesian learning model in my conceptual framework, the rate of belief updating depends on the uncertainty of prior beliefs. By this logic, policymakers who were less confident about their priors should update more as they place greater weight on the information signal received. Panel (a) shows the learning rate for policymakers with prior variance above and below the sample mean. Consistent with a Bayesian learning model, I find that

policymakers with weak priors are significantly more likely to update their beliefs towards the information they received.

I also find that policymakers are significantly more likely to update their beliefs if the information they received matches with their most preferred source of information. The result is captured in panel (b), which shows that the learning rate was 8.5 percentage points higher for those who were randomly assigned to the information they preferred to see. This finding underscores the importance of generating genuine interest and demand among policymakers to learn about research evidence if we hope to change their prior beliefs.

Given the large effect of the researcher-plus information group, I examine heterogeneity in treatment effects for this particular group. Specifically, I test whether the treatment effects vary by scientific reasoning ability, context accuracy, and role, in order to understand if certain types of policymakers are more likely to update their beliefs. As shown in panels (c) through (e), the differences in slopes between these policymaker characteristics are not statistically significant. These results suggest that when research evidence is presented in an accessible way to policymakers, they are likely to be widely incorporated into the beliefs of policymakers.

5.4 Follow-up measures

5.4.1 Policy recommendation

In the obfuscated follow-up survey, respondents are given an opportunity to provide policy recommendations to their local education agency. This policy recommendation task is designed to partially address concerns that survey experiments do not capture consequential behaviors, by increasing the real-world stakes associated with taking the survey. Figure 6 summarizes how policymakers ranked the policy issue of charter schools across the different information treatment arms. Since the information treatment described a context where the student population was 12% white and 84% eligible for free- or reduced-price lunch, I only expect the information to be relevant to the policy recommendations made by those who work in relatively similar contexts. To measure similarity, I split the sample to those whose local jurisdictions have values of % white that are below the sample mean (<27%) and values of % FRPL that are above the sample mean (>54%).

Figure 6 shows that policymakers who received the researcher-plus information and work in relatively similar contexts are slightly more likely to rank charter schools as a top-four priority issue relative to other policymakers. I formally test whether these differences across information groups are meaningful by regressing the rank of charter school policy on information assignment using an ordinal logistic regression as follows:

$$log\frac{P(Y \leq j)}{Y > j} = \beta_{j0} - \eta_1 P - \eta_2 R - \eta_3 Rplus - \eta_4 sim - \eta_5 Rplus \times sim \tag{7}$$

where j is the rank, re-scaled as 5 being the highest priority issue and 1 as the lowest priority issue for ease of interpretation. The predictors in the model are indicator variables for assignment to the policymaker treatment (P), the researcher treatment (R), the researcher-plus treatment ($Rplus$), and similarity to the context described in the experiment (sim). The coefficient of interest is η_5, which is the difference in rank order of the researcher-plus treatment relative to the control group, for policymakers working in similar contexts.

The results of equation (7) are presented in Table 5 column 1. The difference in policy ranking between researcher-plus recipients and the control group for those in similar contexts is small ($\eta_5 = 0.132$) and not statistically significant. Columns 2 through 5 in Table 5 estimates equation (7) without the context similarity and interaction terms, for the four other policy issues in the recommendation process. Since the survey experiment did not mention any of these other policy issues, we expect no difference in how policymakers rank these other policy issues by treatment assignment if the experiment worked as intended. The results presented in Table 5 confirm this expectation. Thus, the information treatment has large effects on views about the effectiveness of urban charter schools but does not change real-world policy recommendations made by policymakers.

5.4.2 Text analysis

Finally, I examine the open-ended text response in the follow-up survey. Respondents were asked to explain what informs their policy views about charter school effectiveness. The key advantage with the open-ended question is that respondents are not primed by available answer choices, allowing me to elicit the immediate reason for their policy views.

I analyzed these open-ended responses using text analysis. To capture whether respondents refer to research evidence, the following seed words were used: "research", "study", "evidence", "experiment", and "lottery". All of the relevant synonyms to these seed words were identified in www.thesaurus.com. I combined the seed stems and synonym stems to generate a binary variable for whether policymakers used a research-related stem in their open-ended response. Then, I regressed this binary variable on information treatment assignment to estimate the effect of the treatment on the use of research-related stems. These stem words and text analysis procedure were pre-specified in my pre-analysis plan to eliminate potential degrees of freedom for analysis.

Figure 7 displays the results from the text analysis. The bars display the means for each group and the p-value of the regression coefficients are displayed as brackets above the bars. In the absence of information (control group), 18% of policymakers used research-related stem words when describing what informs their policy views about charters. In contrast, 34.6% of policymakers in the researcher-plus group referred to

research. Notably, the researcher-plus group was significantly more likely to mention research than the researcher group. This finding is consistent with the learning rate results from the belief elicitation task, which showed significantly larger effects of the researcher-plus treatment than the researcher treatment. Taken together, these results suggest that providing policymakers with impact estimates of an education policy is unlikely to change their beliefs and policy views unless they are also provided with an accessible explanation of how the research was conducted.

6 Discussion and Conclusion

The rapid growth of impact evaluations in education over the last two decades has created a rich pool of evidence about what works in education. Policymakers are increasingly expected to use research evidence, yet we know little about their mental models when making evidence-informed decisions. I conducted survey experiments on over 2,000 policymakers working in state and local educational agencies in the United States to study policymakers' preferences for research evidence and how they update their beliefs in response to new information.

My results suggest that policymakers have clear preferences for research evidence. They prefer studies with larger samples, multiple sites, and those conducted in similar settings as their own jurisdiction. This finding has important implications for the production of research in education. Novel ideas are often rewarded in academic publishing, but my results show clear demand for research that tests the efficacy of policies in different contexts, as well as for replication studies that examine the effects of programs at a much larger scale.

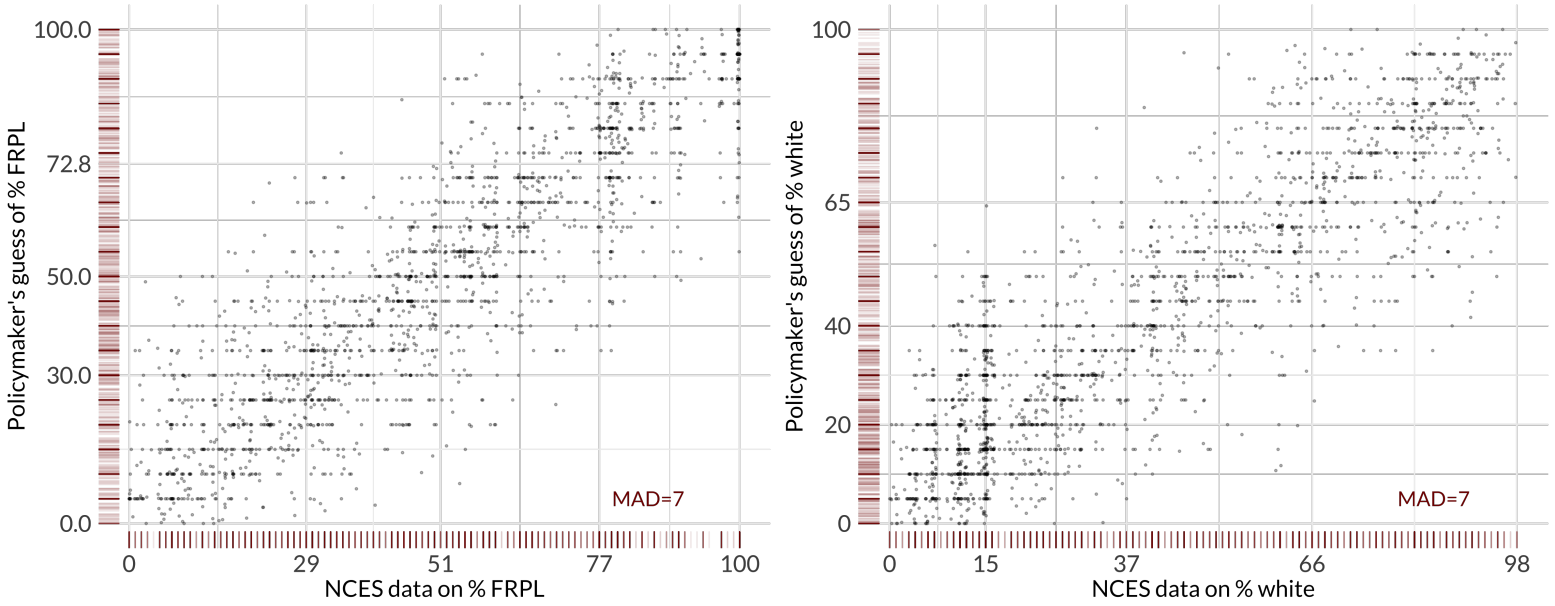
At the same time, policymakers do not show preferences for the design of research studies. They are likely to place equivalent weight to experimental and correlational studies. This result is surprising for two reasons. First, from a perspective of causal inference, experimental studies are less susceptible to threats of internal validity than observational studies. All else equal, experiments are the preferred research design if the goal is to advance policies shown to improve educational outcomes. Second, policymakers in the U.S. are offered federal guidance under ESSA that experiments offer the strongest evidence for causal claims. I show that the average policymaker's indifference towards research design is partly explained by their low scientific reasoning skills. This finding is important because it underscores the importance of training future education leaders to become critical consumers of scientific research.

Finally, I show that policymakers update their beliefs about the effectiveness of education policies when presented with research evidence. These effects are large, persistent, and economically significant only when the research evidence includes a brief, accessible description of the research design. This finding has

important implications for scientific communication. Too often, research is communicated to policymakers as headlines of impact estimates with very little exposition about its research design. My results suggest that policymakers are significantly more likely to incorporate research evidence in their decision-making process if they are able to follow and understand how these estimates were derived in the research presented to them.

Future work might further probe how to ensure that research evidence is actually used to inform policy decisions. In my study, providing policymakers with research evidence changed their beliefs about the effectiveness of an education policy in a belief elicitation task, but it did not change their policy recommendations in a higher-stakes task with possible real-world consequences. While external validity of research often focuses on the generalizability of a policy to the broader population, the relevant challenge for policymakers is the applicability of evidence to their specific context. We need to better understand how policymakers bridge this gap between the best evidence available from other contexts and the idiosyncrasies of their own context, when advancing policies to help improve education.

7 Tables & Figures

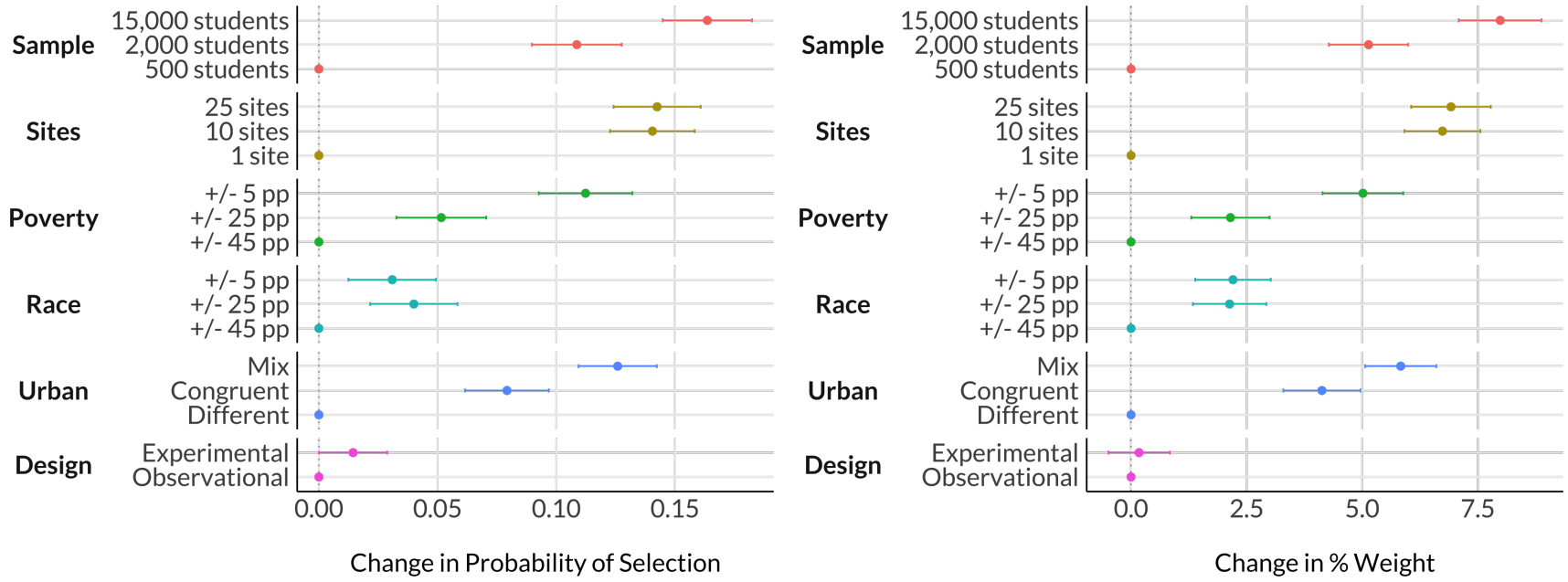


(a) Percent of students eligible for FRPL

(b) Percent of white students

Figure 1: Context accuracy of policymakers

Note: This figure summarizes the context accuracy of policymakers. Each plot is an individual policymaker (N=2,079). In each figure, the x-axis is the statistic reported by the National Center for Education Statistics (NCES) Common Core of Data in 2020 and the y-axis is the policymaker's guess of that statistic. (a) is the percent of students eligible for free- or reduce-price lunch (FRPL) and (b) is the percent of white students. The red markers along the axis display marginal densities of the variables, with labels indicating the 25th, 50th, and 75th percentiles of the variables. A perfectly accurate policymaker would be plotted along a straight, 45-degree line. The median absolute deviation between NCES data and policymakers' guess is reported in the bottom right of each figure.



(a) Forced choice outcome (binary)

(b) Rating outcome (continuous)

Figure 2: Policymaker preference for research studies

Note: This figure presents the estimates of the effects of the randomly assigned study attributes on the probability of being selected to inform policymakers' decisions (a) and the percent weight on policymakers' decisions (b). Estimates are based on the regression specification in equation 1. Standard errors are clustered at the respondent level. Bars represent 95% confidence intervals. The reference category for each attribute is denoted as a point estimate at 0.

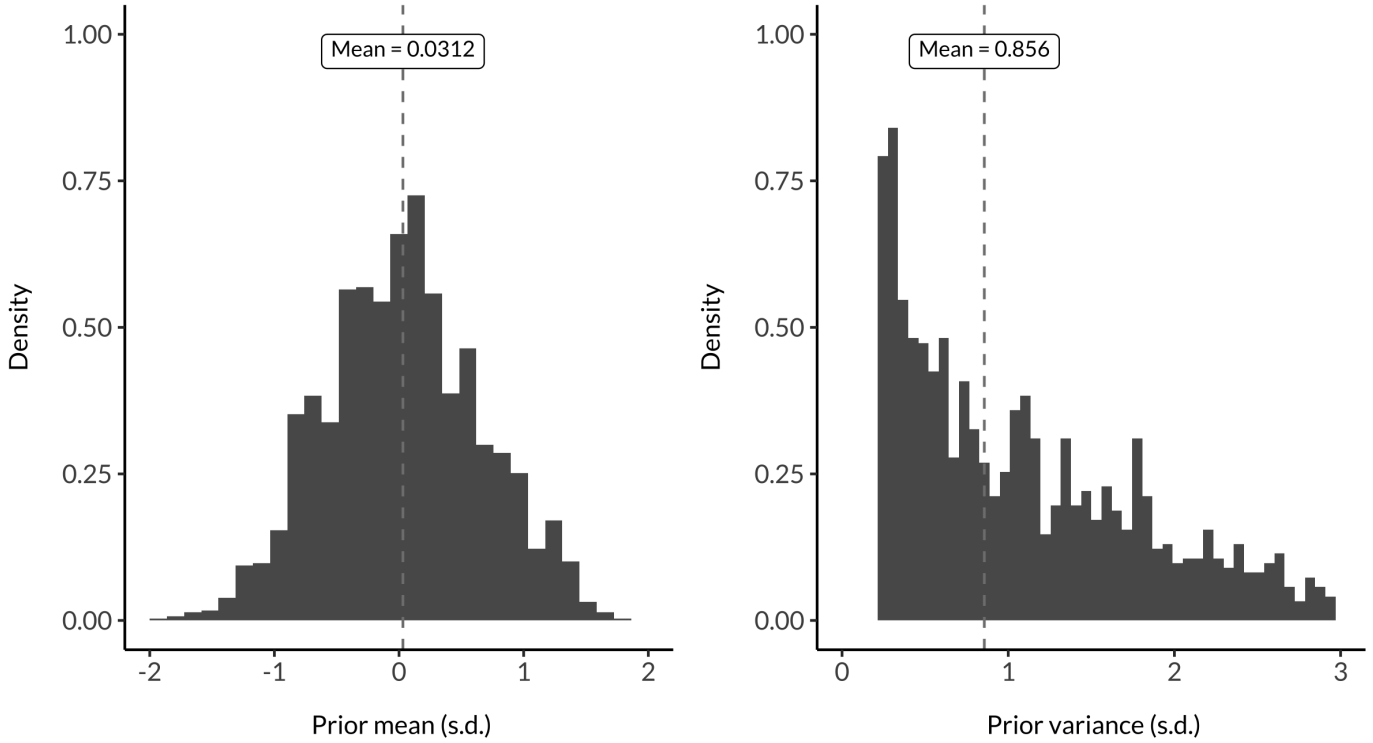
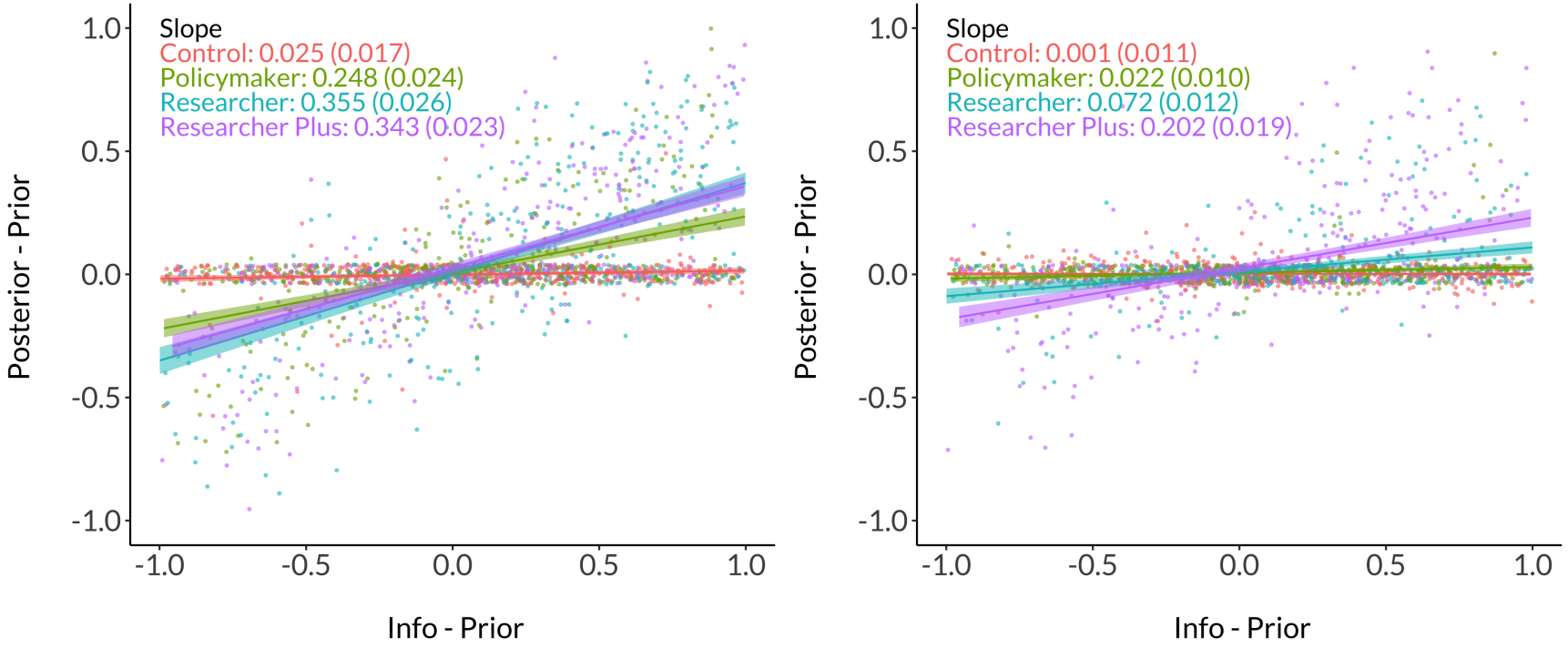


Figure 3: Policymakers' prior beliefs

Note: This figure presents the mean and variance of policymakers' predictions about the effect of urban charter schools on student achievement. These responses were measured at the beginning of the survey. Policymakers reported support points and probabilities associated with each support point. Sample size is 2,079 policymakers.

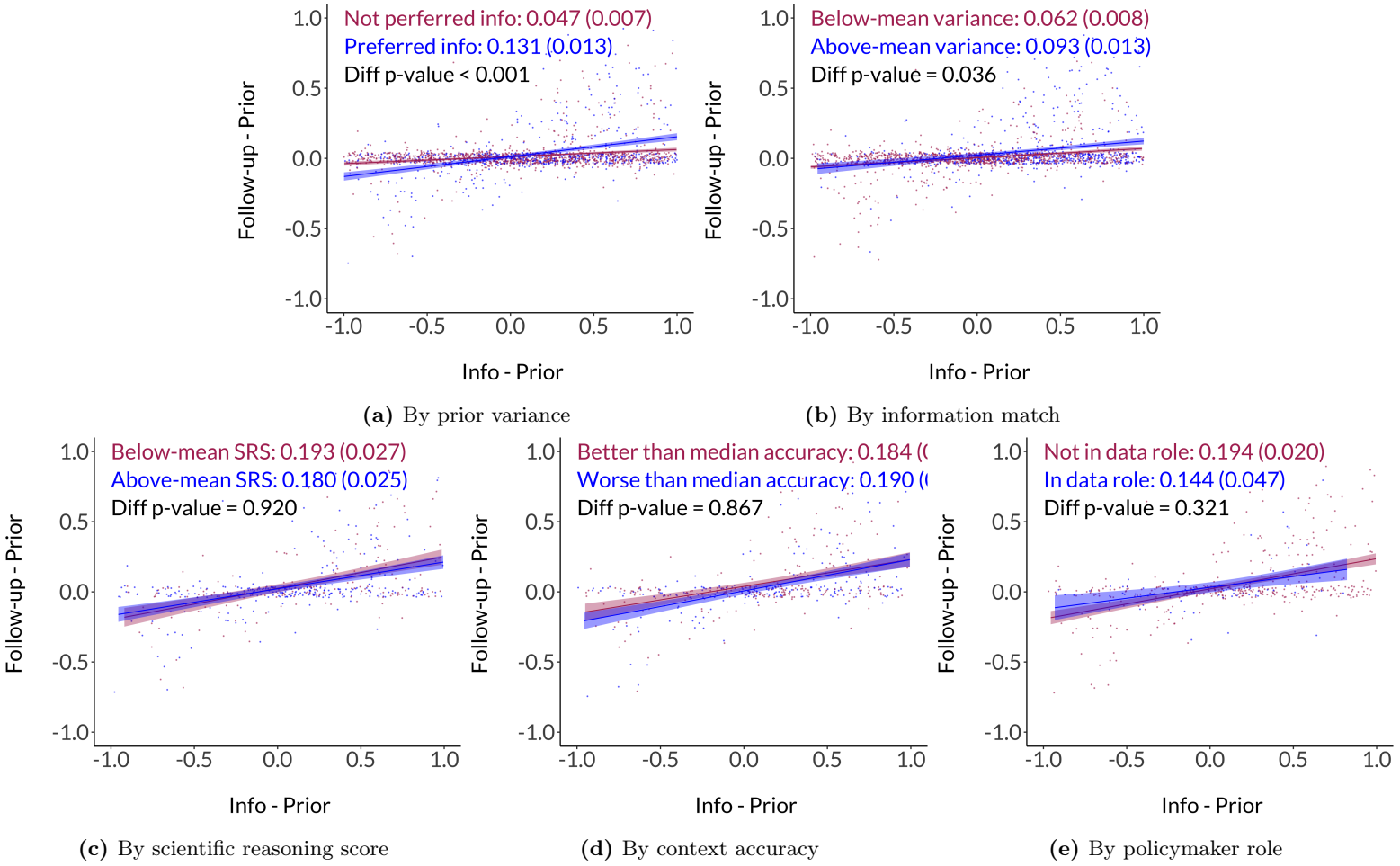


(a) Posterior measured at end of main survey

(b) Posterior measured at follow-up survey

Figure 4: Learning rate of policymakers

Note: This figure presents the learning rates of policymakers estimated using equation 6 from Section 5.3. Each point in the scatter plot is an individual policymaker (N=2,002). The y-axis captures belief update; panel (a)uses the difference between the posterior belief measured at the end of the main survey and the prior belief, and panel (b) uses the difference between the posterior belief measured at the follow-up survey and the prior belief. The x-axis is the signal gap, measured as the difference between the information value shown in the experiment and the prior belief. The regression line includes 95% confidence intervals. Slopes and robust standard errors in parentheses are displayed for each group in the top-left corner.



(a) By prior variance **(b)** By information match

(c) By scientific reasoning score **(d)** By context accuracy **(e)** By policymaker role

Figure 5: Heterogeneity of learning rate of policymakers

Notes: This figure presents the learning rates of policymakers by different covariates. Panel (a) separates policymakers by whether the variance of their prior beliefs was above or below the median. Panel (b) separates by those who were and were not randomly assigned to the information they preferred to see. Panel (c) separates by whether the scientific reasoning score was above ore below the sample mean. Panel (d) separates by whether policymakers' absolute deviation between the NCES data and their guess for %white and %FRPL was above or below the sample median. Panel (e) separates by whether the primary job task of the policymaker includes data and research related activities.

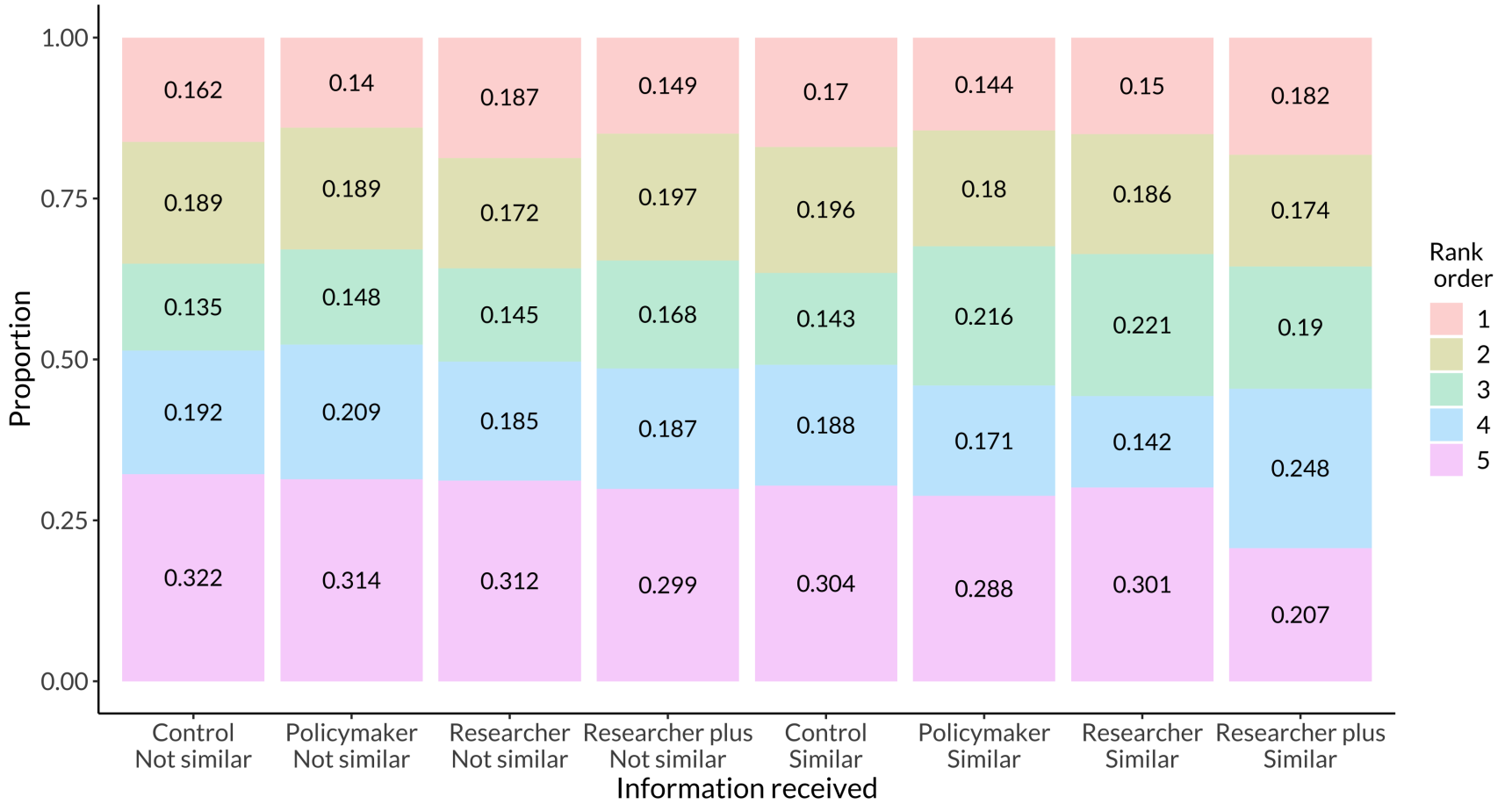
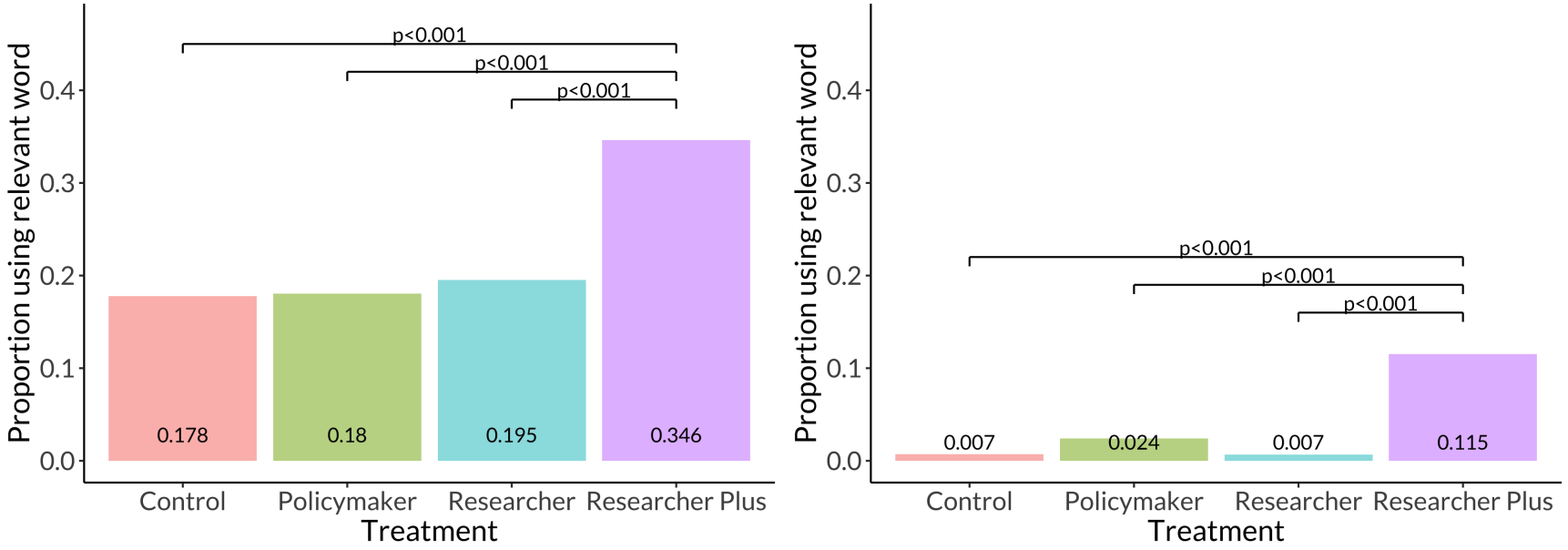


Figure 6: Ranking of charter school expansion in policy recommendation

Note: This figure presents the rank order assigned by policymakers on the policy issue of whether to expand or not expand charter schools in their local jurisdiction. The rank order values are 1 (most important) and 5 (least important). The x-axis labels "Control", "Policymaker", "Researcher" and "Researcher plus" refer to the type of information received in the experiment. The x-axis labels "Not similar" and "Similar" refer to whether the local jurisdiction of the policymaker is similar to the context described in the information experiment: 12% white and 84% eligible for free- or reduced-priced lunch (FRPL). Specifically, the similarity is measured as having values of % white below the sample mean ($< 27\%$) and % FRPL above the sample mean ($> 54\%$).



(a) Research-related seed words

(b) Experiment-related seed words

Figure 7: Response to open-ended question in follow-up survey

Note: This figure presents the results of the open-ended text response in the follow-up survey, which was administered six weeks after the information experiment. Policymakers were asked, "What has informed your policy views about the effectiveness of charters at improving student achievement?" Panel (a) shows the proportion of respondents who used any of the following seed words, by information treatment assignment: "research", "study", "evidence", "experiment", and "lottery" and the synonyms of these seed stems. Panel (b) shows the proportion of respondents who used seed words related to experiments, by information treatment assignment: "experiment" and "lottery". The seed and synonym stems were specified in the study's pre-analysis plan. The figures display 95% confidence intervals and the p-values for tests of equality of means across the conditions.

Table 1: Description of study sample and comparison sample

	Study Sample (%)	American Teacher Panel (%)	American School Leaders Panel (%)
Panel A. Roles			
District Leader	43.2	–	–
Other District Administrator	35.4	–	–
State Admin	14.7	–	–
State Leader	6.6	–	–
Primary job task includes data/research	15.3	–	–
Panel B. Demographics			
Female	79.5	76.5	51.4
White	62.9	83.1	80.1
African American/African/Black	14.0	7.1	12.8
Hispanic/Latino	10.3	7.8	7.8
Asian American/Asian	6.7	2.7	1.7
Panel C. Setting			
Urban	56.4	28.7	25.7
Suburban/Town	33.1	51.0	45.9
Rural	10.4	20.4	28.4
% free- or reduced-price lunch students	51.7	52.25	53.89
% white students	41.5	50.43	53.89

Note: This table reports summary statistics on education policymakers from my study sample, and a comparison sample of nationally representative educators from the 2018 American Teacher Panel (ATP) and the 2018 American School Leader Panel (ASLP). My study sample includes N=2,079 education policymakers, the ATP sample includes N=15,719 K-12 public school teachers, and the ASLP sample includes N=3,540 school leaders.

Table 2: Item-level responses of Scientific Reasoning Scale (SRS)

	Study Sample	Drummond & Fischhoff (2015) Study 2	Drummond & Fischhoff (2015) Study 3
1. Blind/double blind	0.76	0.53	0.61
2. Causality	0.64	0.54	0.58
3. Confounding variables	0.70	0.76	0.79
4. Construct validity	0.71	0.56	0.61
5. Control group	0.80	0.76	0.77
6. Ecological validity	0.75	0.68	0.67
7. History	0.72	0.69	0.72
8. Maturation	0.56	0.66	0.68
9. Random assignment to condition	0.76	0.64	0.65
10. Reliability	0.63	0.49	0.51
11. Response bias	0.69	0.35	0.45
Total items correct (out of 11)	7.7	6.6	7.0

Note: This table reports the proportion of correct responses for each item on the Scientific Reasoning Scale (SRS) from my study sample, and a comparison sample from Study 2 and Study 3 in Drummond & Fischhoff (2015). My study sample includes N=2,079 education policymakers, Study 2 includes N=274 American adults from MTurk, and Study 3 includes N=295 American adults from MTurk.

Table 3: Heterogeneity of policymaker preferences

	Interaction of attributes with:		
	Scientific Reasoning Scale	Context Accuracy	Data Role
	(1)	(2)	(3)
Sample (ref: 500 students)			
2,000 students	0.003	0.001	−0.030
	(0.006)	(0.001)	(0.025)
15,000 students	0.011	0.0002	−0.010
	(0.006)	(0.001)	(0.027)
Sites (ref: 1 site)			
10 sites	−0.002	0.0004	0.024
	(0.006)	(0.001)	(0.025)
25 sites	0.009	−0.0003	0.003
	(0.006)	(0.001)	(0.025)
Poverty (ref: +/- 45 pp)			
+/- 25 pp	0.002	0.0003	0.017
	(0.006)	(0.001)	(0.026)
+/- 5 pp	0.010	0.001	0.028
	(0.006)	(0.001)	(0.026)
Race (ref: +/- 45 pp)			
+/- 25 pp	0.002	0.00001	−0.001
	(0.006)	(0.001)	(0.024)
+/- 5 pp	0.008	0.001	0.015
	(0.006)	(0.001)	(0.025)
Urban (ref: Different)			
Mix	0.005	−0.0002	−0.019
	(0.005)	(0.001)	(0.023)
Congruent	0.003	−0.0001	−0.013
	(0.006)	(0.001)	(0.024)
Design (ref: Observational)			
Experimental	0.013**	0.001	0.040
	(0.005)	(0.001)	(0.021)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: This table presents the interaction terms of the attributes with Scientific Reasoning Scale (column 1), context accuracy (column 2), and data role (column 3) in a regression of study selection on study attributes. SRS is an 11-item measure; higher values indicate better scientific reasoning. Context accuracy is the absolute deviation between NCES data and policymakers' guess of the data; higher values indicate less context accuracy. Data role is an indicator of whether the primary job task includes research or data-related activities; a value of 1 equals yes. Robust standard errors in parentheses are clustered at the respondent level. The reference category for each attribute is denoted in brackets.

Table 4: Information preference of policymakers

	Information most preferred:		
	Researcher forecast	Policymaker forecast	No information
	(1)	(2)	(3)
Scientific reasoning scale	0.016*	0.009***	-0.025***
	(0.007)	(0.037)	(0.007)
Context accuracy	0.001	-0.001	0.0001
	(0.001)	(0.001)	(0.001)
Data Role (0/1)	0.017	0.016	-0.033**
	(0.03)	(0.029)	(0.012)
Female (0/1)	-0.045	0.037	0.008
	(0.026)	(0.025)	(0.013)
White (0/1)	-0.024	0.028	-0.004
	(0.022)	(0.021)	(0.011)
Urban (0/1)	0.009	-0.015	0.005
	(0.022)	(0.021)	(0.011)
Variance of prior belief	0.038***	0.016**	-0.055***
	(0.011)	(0.011)	(0.011)
Mean	0.597	0.340	0.063

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: This table presents the bivariate relationship between the information most preferred (column) and characteristics of policymakers (row). Robust standard errors in parentheses. N=2,002 for each cell.

Table 5: Information preference of policymakers

	Policy recommendation				
	Charter schools	School financing	Accountability standards	Virtual learning	Flexible teacher licensure
	(1)	(2)	(3)	(4)	(5)
Reference: No information					
Policymaker	-0.044	-0.0001	0.016	0.045	0.07
	(0.127)	(0.114)	(0.112)	(0.109)	(0.111)
Researcher	0.072	-0.083	0.084	0.091	-0.003
	(0.131)	(0.111)	(0.113)	(0.110)	(0.113)
Researcher-plus	0.048	0.001	0.097	0.127	0.046
	(0.125)	(0.110)	(0.110)	(0.108)	(0.111)
Similar	0.079	–	–	–	–
	(0.195)	–	–	–	–
Policymaker x Similar	0.029	–	–	–	–
	(0.266)	–	–	–	–
Researcher x Similar context	-0.069	–	–	–	–
	(0.271)	–	–	–	–
Researcher x Similar context	0.132	–	–	–	–
	(0.257)	–	–	–	–

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: This table presents results from the ordinal logistic regressions described in section 5.4. The column indicates the policy issue ranked by policymakers in the follow-up experiment. The rank order values are 1 (least important) and 5 (most important). N=2,002 for each column.

References

Altig, D., Barrero, J. M., Bloom, N., Davis, S. J., Meyer, B., & Parker, N. (2020). Surveying business uncertainty. *Journal of Econometrics*.

Angrist, J. D., Pathak, P. A., & Walters, C. R. (2013). Explaining charter school effectiveness. *American Economic Journal: Applied Economics*, *5*(4), 1–27.

Brighouse, H., Ladd, H. F., Loeb, S., & Swift, A. (2018). *Educational goods: Values, evidence, and decision-making*. University of Chicago Press.

Bursztyn, L., Haaland, I. K., Rao, A., & Roth, C. P. (2020). *Disguising prejudice: Popular rationales as excuses for intolerant expression* (Tech. Rep.). National Bureau of Economic Research.

Center for Research on Education Outcomes. (2009). Multiple choice: Charter school performance in 16 states. *Technical Report*.

Center for Research on Education Outcomes. (2015). Urban charter school study. *Technical Report*.

Chabrier, J., Cohodes, S., & Oreopoulos, P. (2016). What can we learn from charter school lotteries? *Journal of Economic Perspectives*, *30*(3), 57–84.

Cheng, A., Henderson, M., Peterson, P. E., & West, M. R. (2019). Public support climbs for teacher pay, school expenditures, charter schools, and universal vouchers: Results from the 2018 ednext poll. *Education Next*, *19*(1), 8–27.

Coburn, C. E., Honig, M. I., & Stein, M. K. (2009). What's the evidence on districts' use of evidence. *The role of research in educational improvement*, 67–87.

Coburn, C. E., & Talbert, J. E. (2006). Conceptions of evidence use in school districts: Mapping the terrain. *American journal of Education*, *112*(4), 469–495.

Cohodes, S., Setren, E. M., & Walters, C. R. (2021). Can successful schools replicate? scaling up boston's charter school sector. *American Economic Journal: Economic Policy*, *13*(1), 138–67.

Cohodes, S., Setren, E. M., Walters, C. R., Angrist, J. D., Pathak, P. A., et al. (2013). *Charter school demand and effectiveness: A boston update*. Massachusetts Department of Elementary and Secondary Education.

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, MA.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of economic literature*, *48*(2), 424–55.

Drummond, C., & Fischhoff, B. (2017). Development and validation of the scientific reasoning scale. *Journal of Behavioral Decision Making*, *30*(1), 26–38.

Every student succeeds act, pub.l. 114–95. (2015).

Furgeson, J., Gill, B., Haimson, J., Killewald, A., McCullough, M., Nichols-Barrer, I., ... others (2012). Charter-school management organizations: Diverse strategies and diverse student impacts. *Mathematica Policy Research, Inc.*.

Fuster, A., Perez-Truglia, R., Wiederholt, M., & Zafar, B. (2018). Expectations with endogenous information acquisition: An experimental investigation. *The Review of Economics and Statistics*, 1–54.

Gleason, P., Clark, M., Tuttle, C. C., & Dwoyer, E. (2010). The evaluation of charter school impacts: Final report. ncee 2010-4029. *National Center for Education Evaluation and Regional Assistance*.

Gordon, N., & Conaway, C. (2020). *Common-sense evidence: The education leader?s guide to using data and research*. Harvard Education Press.

Grissom, J. A., & Andersen, S. (2012). Why superintendents turn over. *American Educational Research Journal*, *49*(6), 1146–1180.

Haaland, I., Roth, C., & Wohlfart, J. (2020). Designing information provision experiments.

Hainmueller, J., Hangartner, D., & Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences*, *112*(8), 2395–2400.

Hightower, A. M. (2002). *School districts and instructional renewal* (Vol. 8). Teachers College Press.

Hill, H., & Briggs, D. (2020). Education leaders? knowledge of causal research design: A measurement challenge. *Annenberg EdWorking Papers*.

Hjort, J., Moreira, D., Rao, G., & Santini, J. F. (2021). How research affects policy: Experimental evidence from 2,150 brazilian municipalities. *American Economic Review*, *111*(5), 1442–80.

Honig, M. I., & Coburn, C. (2008). Evidence-based decision making in school district central offices: Toward a policy and research agenda. *Educational Policy*, *22*(4), 578–608.

Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation*, *30*(3), 377–410.

Kirst, M. W. (2007). Politics of charter schools: Competing national advocacy coalitions meet local politics. *Peabody Journal of Education*, *82*(2-3), 184–203.

Liaqat, A. (2019). *No representation without information: Politician responsiveness to citizen preferences* (Tech. Rep.). Working Paper.

Maestas, N., Mullen, K. J., Powell, D., Von Wachter, T., & Wenger, J. B. (2018). *The value of working conditions in the united states and implications for the structure of wages* (Tech. Rep.). National Bureau of Economic Research.

Mas, A., & Pallais, A. (2017). Valuing alternative work arrangements. *American Economic Review*, *107*(12), 3722–59.

National Academies of Sciences, E., Medicine, et al. (2017). *Communicating science effectively: A research agenda*. National Academies Press.

Penuel, W. R., Briggs, D. C., Davidson, K. L., Herlihy, C., Sherer, D., Hill, H. C., ... Allen, A.-R. (2017). How school and district leaders access, perceive, and use research. *AERA Open*, *3*(2), 2332858417705370.

Rogger, D., & Somani, R. (2018). Hierarchy and information. *World Bank Policy Research Working Paper*(8644).

Roth, C., & Wohlfart, J. (2020). How do expectations about the macroeconomy affect personal expectations and behavior? *Review of Economics and Statistics*, *102*(4), 731–748.

Schalet, A. T., Tropp, L. R., & Troy, L. M. (2020). Making research usable beyond academic circles: A relational model of public engagement. *Analyses of Social Issues and Public Policy*, *20*(1), 336–356.

Spillane, J. P. (1996). School districts matter: Local educational authorities and state instructional policy. *Educational Policy*, *10*(1), 63–87.

Spillane, J. P. (2000). Cognition and policy implementation: District policymakers and the reform of mathematics education. *Cognition and instruction*, *18*(2), 141–179.

Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy implementation and cognition: Reframing and refocusing implementation research. *Review of educational research*, *72*(3), 387–431.

Stantcheva, S. (2020). *Understanding economic policies: What do people know and how can they learn* (Tech. Rep.). Working paper.

Stuart, E. A., Bell, S. H., Ebnesajjad, C., Olsen, R. B., & Orr, L. L. (2017). Characteristics of school districts that participate in rigorous national educational evaluations. *Journal of research on educational effectiveness*, *10*(1), 168–206.

Tipton, E. (2014). How generalizable is your experiment? an index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, *39*(6), 478–501.

Tipton, E., Fellers, L., Caverly, S., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Ruiz de Castilla, V. (2016). Site selection in experiments: An assessment of site recruitment and generalizability in two scale-up studies. *Journal of Research on Educational Effectiveness*, *9*(sup1), 209–228.

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, *47*(8), 516–524.

Vivalt, E., & Coville, A. (2020). How do policymakers update their beliefs? *Working paper*.

Weiss, C., Bucuvalas, M. J., & Bucuvalas, M. J. (1980). *Social science research and decision-making*. Columbia University Press.

Weiss, M., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, *33*(3), 778–808.

Weiss, M., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, *10*(4), 843–876.

Wiswall, M., & Zafar, B. (2018). Preference for the workplace, investment in human capital, and gender. *The Quarterly Journal of Economics*, *133*(1), 457–507.

Yong, E. (2012). Replication studies: Bad copy. *Nature News*, *485*(7398), 298.

Zimmer, R., Gill, B., Booker, K., Lavertu, S., & Witte, J. (2012). Examining charter student achievement effects across seven states. *Economics of Education Review*, *31*(2), 213–224.

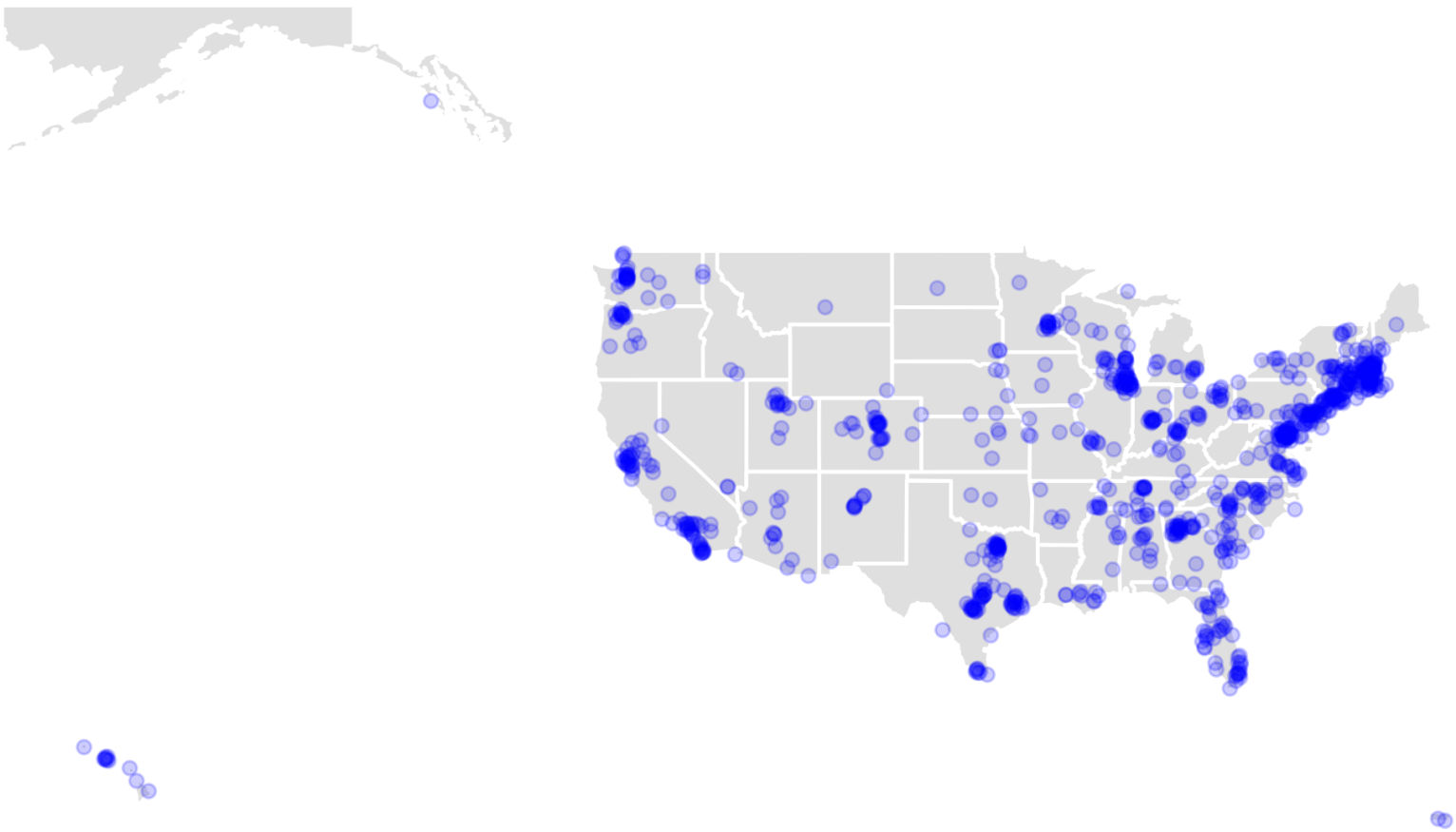


Figure A1: Geographic location of policymakers in study sample

Note: This figure shows the geographic location of the education agency that employs the policymakers in my study sample (N=2,079). The ZIP codes of the education agencies come from administrative records described in Section 3. The policymakers work in 49 states, Washington, D.C. and Puerto Rico.

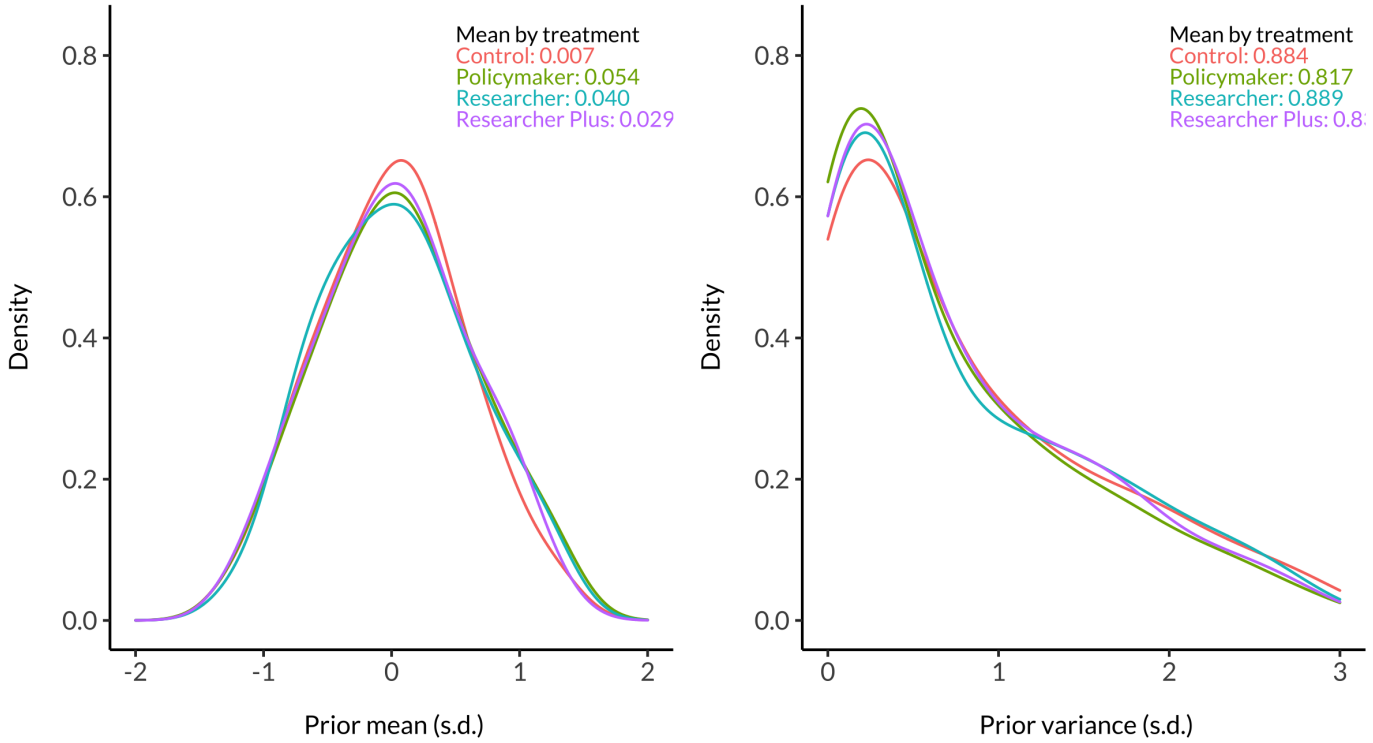


Figure A2: Prior mean and variance by treatment assignment in information experiment

Note: This figure presents the mean and variance of policymakers' predictions about the effect of urban charter schools on student achievement, grouped by information treatment. These responses were measured at the beginning of the survey. Policymakers reported support points and probabilities associated with each support point. Sample size is 2,079 policymakers.

Table A1: Policymaker preferences by model specifications

Dependent variable:	Forced choice		Percent weight
Model specification:	Linear probability	Logit	Linear regression
	(1)	(2)	(3)
Sample (ref: 500 students)			
2,000 students	0.109***	0.459***	5.137***
	(0.010)	(0.041)	(0.437)
15,000 students	0.164***	0.689***	7.983***
	(0.010)	(0.042)	(0.457)
Sites (ref: 1 site)			
10 sites	0.140***	0.592***	6.735***
	(0.009)	(0.039)	(0.419)
25 sites	0.142***	0.598***	6.919***
	(0.009)	(0.040)	(0.440)
Poverty (ref: +/- 45 pp)			
+/- 25 pp	0.052***	0.218***	2.150***
	(0.010)	(0.041)	(0.432)
+/- 5 pp	0.112***	0.474***	5.016***
	(0.010)	(0.043)	(0.446)
Race (ref: +/- 45 pp)			
+/- 25 pp	0.040***	0.169***	2.132***
	(0.009)	(0.040)	(0.405)
+/- 5 pp	0.031**	0.130**	2.204***
	(0.009)	(0.040)	(0.417)
Urban (ref: Different)			
Mix	0.126***	0.523***	5.832***
	(0.008)	(0.035)	(0.393)
Congruent	0.079***	0.328***	4.128***
	(0.009)	(0.037)	(0.426)
Design (ref: Observational)			
Experimental	0.014	0.060	0.174
	(0.007)	(0.031)	(0.339)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: This table presents the estimates of the effects of the randomly assigned study attributes on the probability of being selected to inform policymakers' decisions (forced choice) and the percent weight on policymakers' decisions (percent weight). Estimates are based on the regression specification in equation 1. Each column is the result of a different model specification. Robust standard errors are clustered at the respondent level. The reference category for each attribute is denoted in brackets.

Table A2: Robustness check of policymaker preference for research

	Panel A. Outcome: Forced choice									
	Task number		Order of study		Order of attributes		Order of questions		Placebo	
Attribute	F-statistic	p-value	F-statistic	p-value	F-statistic	p-value	F-statistic	p-value	Coef	(S.E.)
Design	0.49	0.743	0.131	0.718	1.867	0.172	1.186	0.276		
Sample	1.001	0.432	0.002	0.998	2.912	0.054	2.199	0.111		
Sites	0.839	0.568	0.08	0.923	0.394	0.674	0.766	0.465		
Poverty	1.66	0.103	0.597	0.551	1.223	0.294	0.207	0.813		
Race	1.426	0.18	1.167	0.311	0.224	0.799	1.53	0.216		
Urban	1.055	0.392	1.316	0.268	1.831	0.16	0.696	0.499		
Color of column									0.003	(0.007)
	Panel B. Outcome: Percent weight									
	Task number		Order of study		Order of attributes		Order of questions		Placebo	
Attribute	F-statistic	p-value	F-statistic	p-value	F-statistic	p-value	F-statistic	p-value	Coef	(S.E.)
Design	0.296	0.880	0.676	0.411	1.914	0.167	0.035	0.851		
Sample	1.599	0.119	0.087	0.917	1.805	0.165	1.054	0.349		
Sites	1.077	0.376	0.844	0.430	0.151	0.860	0.711	0.491		
Poverty	0.843	0.565	0.081	0.922	0.944	0.389	0.207	0.813		
Race	1.760	0.080	2.271	0.103	0.194	0.824	1.774	0.170		
Urban	0.663	0.725	0.965	0.381	1.175	0.309	0.875	0.417		
Color of column									0.294	(0.320)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table presents robustness checks for the effects of the randomly assigned study attributes on the probability that the study is selected to inform policymakers' decisions (Panel A) and the percent weight of the study on policymakers' decisions (Panel B). Each column is the result of a separate regression (N=20,790). Column 1 interacts study attributes with task number, which can range from task 1 to task 5. Column 2 interacts study attributes with the order of study, which can take on values of study A or study B. Column 3 interacts study attributes with the order of attributes, which can take on values from first to sixth. Column 4 interacts study attributes with the order of questions, which can either be that the forced choice question appeared first or the percent weight question appeared first. Columns 1-4 report the F-statistic and corresponding p-value for a test of whether the effect of the study attribute is equivalent across the interaction term. Column 5 reports the effect of the placebo (the color of the shaded column in the task table) on policymakers' responses. Robust standard errors clustered at the respondent level in parentheses.

Table A3: Balance of baseline characteristics of policymakers in analytic sample

	No information		Peer policymaker		Researcher		Researcher Plus		p-value of diff. in means		
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	(3) - (1)	(5) - (1)	(7) - (1)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Female	0.792	0.406	0.811	0.392	0.779	0.415	0.795	0.404	0.462	0.605	0.931
White	0.619	0.486	0.622	0.485	0.639	0.481	0.631	0.483	0.913	0.511	0.675
Black	0.148	0.355	0.115	0.319	0.124	0.330	0.165	0.372	0.128	0.270	0.446
Hispanic	0.104	0.305	0.113	0.317	0.114	0.318	0.090	0.287	0.645	0.621	0.464
Asian	0.072	0.259	0.078	0.268	0.063	0.243	0.060	0.237	0.783	0.166	0.389
District Leader	0.447	0.498	0.456	0.499	0.404	0.491	0.420	0.494	0.414	0.992	0.450
District Admin	0.355	0.479	0.331	0.471	0.355	0.479	0.378	0.485	0.717	0.131	0.947
State Leader	0.060	0.238	0.068	0.252	0.069	0.254	0.065	0.247	0.613	0.561	0.723
Data Role	0.172	0.377	0.158	0.365	0.118	0.323	0.167	0.373	0.567	0.016	0.843
Urban	0.575	0.495	0.567	0.496	0.576	0.495	0.543	0.499	0.797	0.970	0.309
Rural	0.096	0.295	0.094	0.293	0.114	0.318	0.117	0.322	0.943	0.361	0.271
% FRPL students	52.279	28.083	51.721	29.156	52.402	28.618	50.267	28.197	0.760	0.946	0.254
% white students	42.443	28.201	41.637	28.325	39.726	28.006	42.119	28.060	0.654	0.128	0.854
Scientific reasoning scale	7.701	1.525	7.817	1.608	7.590	1.522	7.701	1.565	0.243	0.254	1.000
Absolute deviation of data and guess of FRPL students	9.775	8.745	9.991	9.043	10.432	9.551	10.083	9.422	0.704	0.259	0.588
Absolute deviation of data and guess of white students	9.675	9.247	10.430	9.776	9.981	9.129	9.954	9.531	0.213	0.600	0.636
Observations	501		487		493		521				

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table summarizes policymaker characteristics collected before the information treatment in the survey experiment for my analytic sample (N-2,002). Columns 1-8 report means and standard deviations for each treatment group. Columns 9-11 report p-values of the mean differences between treatment groups.

Table A4: Balance of baseline characteristics of policymakers that started surveys

	No information		Peer policymaker		Researcher		Researcher Plus		p-value of diff. in means		
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	(3) - (1)	(5) - (1)	(7) - (1)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Attrited (1 = Yes)	0.0649	0.247	0.0737	0.261	0.0722	0.259	0.0843	0.278	0.578	0.641	0.209
Female	0.778	0.416	0.816	0.388	0.769	0.422	0.799	0.401	0.123	0.708	0.386
White	0.629	0.484	0.613	0.487	0.637	0.481	0.642	0.48	0.596	0.774	0.631
Black	0.144	0.352	0.12	0.325	0.121	0.326	0.164	0.37	0.232	0.255	0.362
Hispanic	0.103	0.304	0.11	0.314	0.116	0.32	0.086	0.281	0.676	0.494	0.335
Asian	0.068	0.253	0.077	0.267	0.069	0.253	0.056	0.229	0.793	0.212	0.965
District Leader	0.441	0.497	0.449	0.498	0.404	0.491	0.44	0.497	0.493	0.968	0.747
District Admin	0.357	0.479	0.337	0.473	0.356	0.479	0.366	0.482	0.816	0.181	0.597
State Leader	0.061	0.24	0.068	0.252	0.07	0.256	0.064	0.245	0.644	0.54	0.844
Data Role	0.173	0.379	0.157	0.364	0.121	0.326	0.162	0.369	0.464	0.015	0.616
Urban	0.566	0.496	0.571	0.495	0.569	0.496	0.55	0.498	0.864	0.925	0.586
Rural	0.094	0.292	0.098	0.297	0.108	0.311	0.116	0.321	0.826	0.42	0.211
% FRPL students	52.085	27.995	52.227	29.169	52.765	28.592	50.567	27.908	0.935	0.689	0.358
% white students	42.4	28.088	41.059	28.361	39.72	28.29	42.285	27.992	0.432	0.114	0.945
Scientific reasoning scale	7.721	1.521	7.796	1.596	7.621	1.538	7.693	1.563	0.427	0.278	0.762
Absolute deviation of data and guess of FRPL students	9.974	8.895	10.117	9.262	10.223	9.377	9.993	9.404	0.795	0.65	0.972
Absolute deviation of data and guess of white students	9.723	9.473	10.504	9.807	10.099	9.216	9.968	9.584	0.18	0.504	0.664
Observations	555		543		554		593				

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: This table summarizes policymaker characteristics collected before the information treatment in the survey experiment for the full sample (i.e., anyone who started the survey). Columns 1-8 report means and standard deviations for each treatment group. Columns 9-11 report p-values of the mean differences between treatment groups. Attrition is defined as respondents who did not complete the main survey and/or the follow-up survey.