Do Political Practitioners Care About Policy Experiments?

Andrew J. Clarke, Julia Flam, and Giulia Matteucci*

November 6, 2026

Word count: 8,869

Abstract

Do political practitioners care about policy experiments? We test this question by providing real evidence from randomized controlled trials (RCTs) to a national sample of verified legislators, school board members, legislative staffers, and other politically engaged citizens. More specifically, we test if positive results increase support and if null results reduce support for educational programs. We also measure the impact of an RCT pilot provision on attitudes in a particularly uncertain policy domain (AI and education). We find that evidence clearly matters to political practitioners, but causal claims without reference to methodology (i.e., "vague" evidence) matter just as much as those backed by rigorous RCTs. These findings raise new questions about the mechanisms of policy diffusion in American laboratories of democracy.

*Andrew Clarke (clarkeaj@lafayette.edu) is the corresponding author. This research was approved by the Institutional Review Board of Lafayette College on December 4, 2023 (Proposal Number: AY2324-37). Our pre-analysis plan (AsPredicted #155752) was registered just before our recruitment email was distributed, on December 18, 2023. In preparing this manuscript, the authors occasionally used AI tools (Claude and Overleaf AI) for copyediting of the text and the proofreading and tidying of hand-drafted R scripts; Appendix C was an outlier to this rule, as we relied more heavily on Claude-generated first-draft code for some robustness checks (e.g., randomization inference). This research would not have been possible without the tremendous research assistance of the students in the Lafayette College Gov Lab and funding from the Landis Center for Community Engagement. We are also indebted to Stephanie Chan, Dylan Groves, Adam Biener, and the Lafayette College Department of Economics for helpful feedback and guidance at early and formative stages in this project.

For many Americans, the COVID-19 pandemic highlighted the importance and difficulty of integrating high-quality evidence into the policymaking process. Faced with a deadly, unprecedented public health crisis, elected officials had to interpret new evidence and provide deeply consequential guidance to millions of people on a near-daily basis.[1] Despite the profound uncertainty of the moment, those with decision-making authority in thousands of U.S. governments were surrounded by causal *claims*.

The way that political practitioners understood and responded to these claims mattered. In the ensuing months, some public officials took extraordinary and creative measures to encourage the use of newly developed, life-saving vaccines with a remarkably high standard of evidence. By contrast, others echoed reckless claims of quick COVID "cures" ranging from bleach to heartworm medication for horses. Carefully sorting through causal claims, in this context, could lead Americans to save lives (vaccines) or quite literally poison themselves (bleach, ivermectin) (Geller and Stempel 2020).

This particularly salient example places the value of high-quality evidence into sharp relief, but elected officials face uncertainty amid competing causal claims across nearly every policy domain. Unfortunately, policymakers often lack rigorous evaluations of the programs that they enact into public law. Given these challenges, social scientists have devoted considerable effort in search of more robust causal identification strategies. Among these, researchers now routinely develop coordinated field experiments with governmental and advocacy organizations to better inform policy solutions to some of the world's most pressing problems. For example, the advances in anti-poverty programs are so deeply rooted in randomized controlled trials (RCTs) that three economists (Abhijit Banerjee, Esther Duflo, and Michael Kremer) were awarded the Nobel Prize in economics "for their experimental approach to alleviating global poverty" (Nobel Prize Outreach 2019). At the same time, governments across the United States have developed new internal units dedicated to more rigorous program impact evaluation (e.g., the Lab @ DC or the Office of Evaluation Sciences).

Policymakers interested in the causal estimates of programs likely have more information available to them today than at any point in human history, but the presence of high-quality evidence does not, in itself, guarantee influence over public policy decisions. Policymakers may not be well-equipped to evaluate

[1] Many Americans took the highly unusual step of tuning into regular briefings from state governors and bureaucratic officials to better understand how to effectively protect themselves, their families, and their neighbors. In fact, one state governor was awarded an Emmy after giving 111 daily public briefings that received a total of 59 *million* viewers (International Academy of Television Arts and Sciences 2020). (The award was rescinded within a few months in light of sexual harassment allegations (Associated Press 2021).)

program impact among competing, plausible claims. Alternatively, they may simply prioritize questions of external validity (i.e., has this been tested in my community?) over those of internal validity (i.e., has this been proven to work?). Competing values, political pressure to adhere to ideological positions, and many other factors make the incorporation of evidence far from a foregone conclusion. Executing a high-quality RCT in public policy is a tremendous feat, but if the goal is effective government, research is only one step in a longer, more complicated process (Cairney 2017).

In this research, we consider the importance of evidence in the minds of political practitioners. More specifically, we test whether state lawmakers, state legislative staffers, school board officials, and other politically engaged Americans are more or less likely to support an education policy if presented with experimental evidence. Our work closely connects to recent social science investigations of the role of evidence in public policymaking (Lee 2022; Hjort et al. 2021; Meyer et al. 2019; Baekgaard et al. 2019) and mechanisms of policy diffusion (Gilardi and Wasserfallen 2019; Gilardi, Shipan, and Wüest 2021; Matthews and Boehmke 2026; Volden 2006). We contribute to these bodies of research by carefully evaluating how policymakers receive both positive *and* (the much more common) null results from well-designed policy experiments, alongside a forward-looking RCT pilot proposal.

The paper proceeds as follows. First, we review directly relevant research on the politics of evidence in laboratories of American democracy. Second, we detail the nature of our custom sample of verified policymakers and politically active Americans. We then describe three experiments designed to test perceptions of evidence in three different contexts. Finally, we conclude with a discussion of what we believe to be next steps towards better understanding the relationship between evidence and effective American governance.

The Politics of Policy Experiments

Careful, empirical evaluations of public policies are particularly useful when they are able to establish fair comparisons, paring back the extraordinary number of confounding factors that would otherwise obscure our understanding of what works and does not work. Among the many tools social scientists use to isolate the impact of a program from alternative explanations, randomized controlled trials (RCTs), or experiments, stand in a category all of their own. Well-conducted experiments help us rigorously test causal claims by creating two groups that, on average, differ only by the presence of their randomly assigned program or public policy intervention. Field experiments of this sort provide the analyst with a reasonable

approximation of what *might have happened* if the public policy "treatment" had never been administered, which is particularly valuable in a political context knotted with countless measurable and unmeasurable influences.[2] From this perspective, it is entirely unsurprising that social scientists have increasingly turned to random assignment in the midst of a "credibility revolution" in empirical studies of social phenomena (Angrist and Pischke 2010).

Notably, researchers have been cognizant of these advantages for years, and experiments in the context of social policy, specifically, appear to have dramatically increased from the mid-20th century to the present day (Baron 2018; Hedges and Schauer 2018; Jatteau 2018). To illustrate this point, Greenberg and Shroder (1997) documented 143 known social experiments conducted in the United States over a thirty-year period. This archive of policy experiments in the U.S. is, in itself, expansive; the collection ranges from Heather Ross's 1967 proposal to test a negative income tax to a wide range of RCTs across a dozen policy domains (some still in the field as of 1996). More recently, Connolly, Keenan, and Urbanska (2018) found *over 1,000 unique experiments* between 1980 and 2016 – a staggering count that is even more remarkable when you consider that their search was restricted to education policy experiments published in English.[3]

Policymakers appear to have taken notice. In 1998, Congress appropriated $150 million per year in funding for schools to adopt "proven, comprehensive reform models" in education policy, where proven was defined by experimental methods with standards-based measures (Slavin 2002, 15). In the years that followed, governments around the world made sustained and considerable investments in searching for impactful, low-cost, light-touch interventions (i.e., "nudges") (Benartzi et al. 2017). These efforts were particularly salient in the United Kingdom, but those efforts have spread throughout policymaking corners of the world. The Behavioral Insights Team in North America (BIT NA) alone conducted 73 trials across 30 U.S. cities between 2015 and 2019 (DellaVigna, Kim, and Linos 2024). The development of the Office of Evaluation Sciences (OES) has led to landmark studies centrally located within the U.S. executive branch (OES 2021, OES 2025), and a bipartisan act of Congress recently restructured the ways in which evidence is incorporated into federal policymaking and funding decisions through the Foundations for Evidence-Based Policymaking Act of 2018 (Haskins and Margolis 2014; Shea 2004).

[2] As Stevenson (2023) points out, RCTs have benefits that go beyond the elimination of selection bias and other problems of causal inference. Because RCTs are typically costly studies with reduced researcher degrees of freedom, this approach has the relative advantage of minimizing the incentive to adjust an analysis until the desired results appear.

[3] This count includes trials located across several continents, but over half of the RCTs in their systematic review (543) were conducted in the USA and Canada alone.

Despite the heightened interest in government-facilitated RCTs, many have noted that policy experiments come with their own methodological challenges. For example, random assignment can greatly increase the internal validity of a study, but successful interventions may run into considerable limitations if they are applied in new contexts. Simply put, what works in one community may not work elsewhere, and the reality of adapting promising initiatives to new contexts is deceptively complex (Knox, Hill, and Berlin 2018). Consider a study that fails to replicate an otherwise successful program to increase student learning outcomes. One might reasonably wonder if the preliminary evidence was specific to a local context; yet it is also possible that the true causal effects persist across contexts but ultimately remain undetected due to fidelity issues of implementation. The data from replication studies may bundle both the positive underlying impact of a program with countervailing administrative capacity limitations specific to a new place. As Cartwright and Munro (2010) put it, null findings may thus be "a result of all the contributions, positive and negative, 'added' together, just as when a magnet and gravity act together on a pin that doesn't fall" (p. 262). So, experimental treatments may simply work in one place and fail in another – or they may work across contexts but encounter difficult, bundled treatment or fidelity problems.

Beyond these methodological challenges, policymakers may simply reject the assumed hierarchy of evidence adopted by many advocates for evidence-based policymaking. Other scholars have made this point forcefully, arguing against

> the lazy assumption that elected policymakers are the villains of this piece because they, for example, do not understand RCTs and the need for RCT-driven evaluations, do not recognise a hierarchy of evidence in which the systematic review of RCTs represents the gold standard, and/or are unwilling to overcome ethical dilemmas about who gets to be in/out of the control group of a promising intervention (Cairney and Oliver 2017, p. 9).

Anecdotal accounts from constituents and politically valuable stakeholders may matter to policymakers, even if they fully understand the methodological shortcomings of a non-random set of inputs fraught with selection bias. Public officials may simply balk at the ethics of experimentation or prioritize expressions of values over measured impact.

Policy experiments can also come with political baggage. Elected officials may face tremendous pressure from citizens who would rather not be treated as test subjects by their governments. For example, researchers evaluating the impact of social welfare programs in Florida in the 1990s were compared to both Nazis and the unethical researchers of the infamous Tuskegee Study of syphilis in the 1930s; their plan was labeled as "misguided, wasteful, cold-hearted, and just plain dumb" (Gueron and Rolston 2013, p.

302). Policymakers may agree to support an RCT after considerable deliberation and detailed negotiations – only to be convinced by critical headlines that experiments are not worth the associated political risks.

Even without public pressure, administrators might be staunchly resistant to the idea of random assignment from within the government. Staff with extensive first-hand experience working on a policy problem may express genuine concern over randomly withholding programmatic benefits, rather than allocating by perceived need, for example. Alternatively, those responsible for implementing a program may worry about political and professional blowback from small or null results. As Gueron notes, "It takes courage for political appointees to favor independent studies that measure net impacts" (i.e., average treatment effects) because agreeing to an RCT means you lose some control over the story and, more importantly, the program's "impacts are almost always smaller than outcomes" (Gueron 2002, p. 24).

Taken together, experimental methods may present the best approach to identifying causal effects when controlled interventions are feasible, but there are several reasons that political practitioners may oppose a complete and unconditional investment in policy experimentation. These approaches are costly and slow. They face external validity concerns, and they run into significant political challenges from both the public and other necessary stakeholders in the policymaking process. The necessary incrementalism of policy experimentation may move too slowly for a society in desperate need of bold action, and the accumulation of small interventions may never produce meaningful and lasting changes even if we are comfortable with this sort of incrementalism (Stevenson 2023).

Given these tradeoffs – and the extraordinary rise in social science field experiments – this research article asks a simple set of questions. Do political practitioners care about evidence, and if so, do they view experimental evidence as more important in their assessment of a potential policy? How do they respond to disappointing null results and surprising positive results? Because policymakers seek out information – particularly from experts – as they weigh in on salient policy issues (Fagan and McGee 2022), these questions are vital to our understanding of political representation, lobbying, and policymaking. We turn next to relevant empirical work on policy experiments to introduce and contextualize our expectations.

Expectations

Existing studies define "high-quality evidence" in a variety of defensible – but far from settled – ways. In fact, the term "evidence-based policymaking" is in itself a contested and routinely undefined phrase (Cairney 2017). Some scholars assume social scientific experimentation has an obvious parallel to evidence-

based medicine, but as others have noted, this approach can fail to grapple with particular political considerations, such as the need to legitimize policy choices or prioritize a sense of accountability in systems with free and fair elections (Cairney 2017). Because policymaking is inherently value-driven work (Cairney and Oliver 2017), we do not *assume* a shared hierarchy of evidence among researchers and policymakers. Instead, we investigate perceptions of experimental evidence directly.

The latest and most relevant empirical social science research on perceptions of evidence has varied the nature of research findings presented to political practitioners in many different ways. Some investigate differences in responses to evidence consistent across multiple studies involving large samples (Bergan et al. 2026) or conclusions that reflect a consensus among domain-relevant experts (Lee 2022). Prior work has shown particular interest in "high-quality evidence," with researchers measuring "quality" in a wide variety of ways, including: careful causal identification strategies, robust sample sizes, numerous study sites, variation in the location of the study, larger point estimates, sampling strategies that improve external validity, and/or estimates with smaller confidence intervals (Crawfurd et al. 2025; Nakajima 2021; Pereira et al. 2025; Vivalt, Coville, and Kc 2025; Vivalt and Coville 2023). In some cases, researchers have taken deliberately expansive approaches to understanding perceptions of evidence quality through discrete choice experiments that vary many of the aforementioned attributes in hypothetical policy scenarios. By contrast, we took a decidedly simplified approach; we provided accessible summaries of real-world studies in a sequence of vignettes. We chose to avoid making broad claims around high-quality evidence, broadly defined, and instead aimed to better understand how evidence from randomized controlled trials (RCTs) specifically matters – or does not matter – to policymakers who often hold the authority to implement research findings or fund new pilot programs.[4]

Our first expectation is that policymakers respond to evidence that a program works. Much of the empirical work on legislative politics assumes that information affects policymaking, and a recent field experiment conducted with state legislators shows that one-on-one staff briefings can impact cosponsorship rates in some instances (Zelizer 2018). More recently, Toma and Bell (2024) find that decision aids (e.g., side-by-side comparisons with impact calculators) can meaningfully impact policymakers' perception of information; they conclude that the relevant barrier to evidence-informed policymaking is the complexity

[4]Whenever possible, we also stuck to real claims from existing, well-designed RCTs accompanied by plain-language explanations for why random assignment is useful. The only exception to this is our fourth hypothesis, which is interested in a proposal for a new, untested RCT pilot program. Here we allowed students to generate an authentic proposal, but there was no real-world study to reference by definition.

of decisions these individuals must face. In education policy, specifically, leaders report positive attitudes towards and frequent use of research (Penuel et al. 2017).[5] These studies suggest that policymakers are open to considering clearly presented, low-complexity summaries of evidence.

On the other hand, politicians may interpret information through carefully calibrated professional incentive structures, and it is not clear that a simple policy brief is enough to change beliefs of program effectiveness. In survey experiments with 954 Danish city councilors, for example, Baekgaard et al. (2019) find that providing more evidence increases, rather than reduces, the influence of prior attitudes on public policies. Moreover, Masset et al. (2013) find no evidence that disseminating a policy brief changes prior beliefs.[6]

Two recent studies provide perhaps the most directly relevant findings. First, Lee (2022) provided information on several policy positions (needle exchanges, GMO bans, and rent control) to a sample of 690 local and state policymakers in the U.S.[7] Treated respondents were provided with an additional statement reporting that "scientists from several universities have examined this issue" (Lee 2022, 680). Introducing this information improved belief accuracy and preference congruence (Lee 2022, 682–83).[8] As Lee noted, the treatment was intentionally brief (i.e., a single sentence) and contained no citations, methodological description, or arguments, which presumably biased *against* finding any effects at all.[9] We take this study as a strong case that policymakers respond to the introduction of evidence, all else equal.

Hjort et al. (2021) make perhaps an even stronger case by moving beyond expressed statements of support and measuring actual changes in policy. Working with Brazil's National Confederation of Mu-

[5]Future experiments may also incorporate their finding that research is routinely accessed through professional networks.

[6]However, we share the authors' view that the study's low response rates (1%) and high attrition rates (71%) are worth bearing in mind as we consider their findings.

[7]Policies were selected to reflect high levels of expert consensus and variation in the compatibility of expert perspectives with each party's baseline policy position (as of March 2018).

[8]Effects on belief accuracy ranged from 8–14 percentage points. Lee found more mixed results for preference congruence, with some insignificant results and others in the range of 8–9 percentage points. While partisan differences are not the focus of our research, it is worth noting that Lee found that partisan differences in updating were modest and statistically insignificant.

[9]While we think of Lee (2022) as a closely comparable study, our design departs from Lee in several ways. First, we evaluated different presentations of the same evidence in a three-arm trial. Second, we focused on RCT evidence rather than consensus among domain-relevant experts. Third, we introduced real null results from a well-designed study. Fourth, we evaluated a proposed RCT pilot provision. Fifth, while our sampling strategies are similar, our final sample was considerably larger (2,575 respondents, with roughly 850 per condition). Finally, we did not conduct a within-subject analysis that would allow us to better understand changes in policy views over time.

nicipalities (CNM), whose membership includes over 80% of Brazilian municipalities, the authors ran two experiments. In the first, 764 municipal officials (primarily mayors) from 579 municipalities reported prior beliefs about the effect of an early childhood development program and were then offered the chance to *purchase* the results of one randomly selected RCT using lottery tickets for an expenses-paid trip to the United States. Mayors were willing to pay for research, placed higher value on larger-sample studies, did not distinguish between studies from richer and poorer countries, and updated their beliefs without confirmation bias or asymmetric responses to good and bad news (Hjort et al. 2021, 1444). In the second experiment, 1,818 mayors attending a CNM convention were randomized into two groups, with treated mayors invited to a 45-minute session that introduced experimental evidence on taxpayer reminder letters. Several months later, those attending the treated session were 10 percentage points more likely to have implemented the relevant reminder letters (Hjort et al. 2021, 1445). Taken together, these studies suggest that policymakers will respond to positive results in general and experimental results in particular.

> **Hypothesis 1:** *Political practitioners will increase their support for a policy when provided with positive results.*

Null results pose a harder test, because several findings suggest that policymakers process disappointing evidence differently from good news. In experiments conducted at World Bank and Inter-American Development Bank workshops, Vivalt and Coville (2023) document two biases among policymakers, practitioners, and researchers: asymmetric optimism (updating more on good news than bad) and variance neglect (paying less attention to uncertainty than a Bayesian would). These biases suggest that policymakers may attend more to the *direction* of estimates than to the statistical uncertainty critical to interpreting a null result. Similarly, Pereira et al. (2025) describe a pattern resembling asymmetric optimism among local officials in France, Germany, Italy, Israel, Switzerland, and the United Kingdom. DellaVigna, Kim, and Linos (2024) find "surprisingly limited support for the role of evidence" (p. 2753) in predicting which U.S. cities adopted nudges tested by the Behavioral Insights Team (BIT NA); studies with negative, positive, and positive but insignificant estimates were adopted at about the same rates.[10] By contrast, Briscese and List (2024) present null results from a field experiment on college savings accounts to policymakers recruited from the National Association of State Treasurers' (NAST) annual conference and find that policymakers

[10]The stronger predictor of adoption here was pre-existing communication channels, suggesting researchers should take seriously the organizational capacity for programmatic interventions in planning policy experiments.

revised beliefs and shifted resource allocation preferences in response to this new evidence – but may also have become more averse to future experimentation. Given the mixed results from this set of studies, we set out to directly test if political practitioners respond to null results and reduce support for a program that lacks clear, supportive evidence.

> **Hypothesis 2:** *Political practitioners will decrease their support for a policy when provided with null results.*

Next, we wish to directly investigate the importance policymakers place on experimental evidence (rather than evidence, broadly defined). Nakajima (2021) offers the most directly relevant study in their discrete choice experiment to over 2,000 K-12 educators, district administrators, state administrators, and others with leadership positions in education. Consistent with our first two expectations, they find that policymakers seem interested in research and will respond to accessibly presented evidence. However, Nakajima's (2021) findings show that correlational research and lottery-based studies (i.e., randomized trials) are about equally appealing.[11]

Several other articles have evaluated perceptions of experimental evidence, and, like Nakajima (2021), they often find little support for a shared hierarchy of evidence with RCT-generated data at the top. In survey experiments with 931 senior education officials in 35 low- and middle-income countries, Crawfurd et al. (2025) randomly varied the description of real RCTs and found that revealing that a study was an RCT had no effect on the probability that officials updated toward the true effect.[12] By contrast, the findings from Pereira et al. (2025) show that thousands of local officials across six democracies prefer evidence from randomized impact evaluations – but their preference for a lottery system appears to be driven primarily by a sense of fairness, rather than an emphasis on evidence quality. Elected officials in that study did not seem more likely to be persuaded by causal evidence.

There are some studies, however, that suggest experimental evidence matters for the reasons social scientists might expect. First, Mehmood, Naseer, and Chen (2024a) find that Pakistani deputy ministers randomly assigned to receive intensive causal inference training value quantitative evidence more, commission more RCTs, and are more likely to recommend funding for policies with RCT evidence. In a follow-up study, Mehmood, Naseer, and Chen (2024b) further found evidence that Pakistani tax officers'

[11] As Nakajima (2021) notes, the choice to describe RCTs as studies "based on a lottery" (p. 14) required policymakers to understand that such a system overcame selection bias that threatens observational studies through random assignment.

[12] Officials instead responded to larger samples and to evidence from contexts similar to their own.

econometrics training impacted the likelihood of adopting taxpayer letters – but *only* if they receive a causal (rather than correlational) signal about the program's effectiveness. Finally, in another discrete choice experiment, Vivalt, Coville, and Kc (2025) find evidence that RCTs matter more than observational studies to policymakers and policy practitioners in samples drawn from World Bank and Inter-American Development Bank workshops. Consistent with the debate in this literature, we set out to isolate the impact of preferences for real experimental evidence over causal claims that are identical in every way – except that they do not specify the experimental nature of the research findings.

> **Hypothesis 3:** *Political practitioners will be more responsive to results from randomized controlled trials than results that do not specify a research methodology.*

Our final expectation considers prospective views towards policy experimentation. Prior research suggests that political practitioners may have "experimentation aversion" when confronted with the possibility of a randomized pilot program. For example, Meyer et al. (2019) find that Amazon Mechanical Turk (MTurk) respondents rate policies from nine policy domains as *more inappropriate* if a proposed A/B test is included in the description of the proposal. These higher rates of objection to experiments hold even after exploring conditional effects by a number of subgroup characteristics (e.g., baseline scientific knowledge and preferences towards policies). Heck et al.'s (2020) follow-up study once again finds evidence of experimentation aversion. A majority of MTurk respondents across five experiments rated A/B tests as less appropriate than the average policy rating of policy A or B alone – and high percentages of respondents rated proposals to experimentally evaluate two policies as the worst of three options (A, B, or A/B testing).

Others, however, find greater openness to policy experiments. Dur et al. (2024) find clear support for policy experimentation among a nationally representative sample of Dutch voters. Similarly, Dutch politicians were found to hold positive views towards policy experimentation on average, and these views became more positive still after being informed of favorable public sentiment towards RCTs. As mentioned above, Pereira et al. (2025) found that officials in six democracies all preferred to distribute a valuable package of educational resources to children via a lottery rather than on a first-come, first-served basis; while these preferences were not a product of strong desire for greater causal identification, the preference towards a lottery system in general suggests that randomized pilot programs may not run into the experimentation aversion found in other studies. Finally, Chlond, Goeschl, and Lohse (2025) find that vignette-based measures framing experiments in terms of moral approval may significantly *overstate* ex-

perimentation aversion and understate willingness to support randomized trials. Their findings suggest that policy experimentation is considerably more palatable to individuals when they are provided with the possibility of real-world stakes, when uncertainty around a given policy is higher, and when they are asked which option they themselves would implement if they were the decision-maker.

This study contributes to this literature by directly adding language proposing a randomized pilot program – along with a description of the methodological benefits of conducting an RCT – to a document calling for policymakers to evaluate the costs and benefits of policy with a high degree of uncertainty.

> **Hypothesis 4:** *Political practitioners will be more supportive of policy proposals if the proposal includes a randomized controlled trial pilot provision.*

We test these expectations in a new, national survey of verified political practitioners. The results from these analyses have the potential to further contribute to a flourishing literature on policy diffusion. Success may indeed drive the diffusion of public policy (Volden 2006), but we agree with Gilardi and Wasserfallen (2019) that researchers should continue to investigate the technocratic assumptions associated with policy learning across political boundaries. If policymakers are not learning from successful initiatives, for example, other factors, such as imitation, can lead to the spread of bad policies (Shipan and Volden 2021).

Sample

The population of interest for our study is American practitioners in a position to directly or indirectly influence education policy. In recruiting our sample, we expected our response rate to remain low for several reasons. First, policymakers and other politically influential Americans are difficult to reach; they are often extraordinarily busy, constrained by institutional norms that can discourage communicating in an official capacity, and wary of any written expression of preferences that might invite political and/or professional risk. Second, we had no pre-existing communication with our sampling frame, which meant that our cold email invitation was likely to run into (reasonable) skepticism about our identity and worry about cyber security risks. Finally, we offered no financial compensation for participation in our study. Given these circumstances, we wrote in our pre-analysis plan that we expected "a low response rate (no more than 3%);" our overall response rate at the conclusion of our survey was between 3.2% and 3.5% (Table A1 and Table A2). As context, we compare our sample to Lee (2022) in Appendix Table A3.

Despite these known challenges, we worked with a team of undergraduate research assistants to hand-code publicly available contact information on verified policymakers, their staff, and other highly engaged

Americans. To pre-emptively address the threats to statistical power that would come with a low response rate, we assembled an extremely large sampling frame (80,042) that included several distinct sub-samples.

First, we collected a comprehensive database of every public school board member in the United States as of September 2023, covering over 10,000 school districts and 53,518 potential respondents. Many districts explicitly discourage elected officials from external communication as individual board members. Nevertheless, 1,548 verified elected officials from this subgroup shared their thoughts in this survey, making up over 60% of our final sample.

Next, we assembled a list of verified state lawmakers and state legislative staffers in all 50 states. Email addresses were hand-coded from a professionally curated database (Legistorm) after cross-checking a sample of their contact information against publicly available source websites. This produced 7,012 potential state legislators (3,122 Democratic, 3,822 Republican, 68 not clearly affiliated with either major party) and 7,717 state legislative staffers (4,162 Democratic, 3,260 Republican, 295 not publicly affiliated with either major party). Approximately 19% of our final sample were verified state lawmakers or staffers.

We also incorporated local party leaders and candidates. We obtained contact information on current or former county Republican and Democratic Party leaders in all 50 states, producing 8,597 potential respondents (4,978 Democratic and 3,619 Republican). We added 2,205 candidates from North Carolina's 2022 elections, using the State Board of Elections website and eliminating duplicates (e.g., school board candidates already captured in our first subgroup). About 20% of our final sample were verified party leaders or candidates for office.

Finally, we invited a little under 1,000 members of collegiate partisan organizations (482 College Republicans and 511 College Democrats), a younger demographic of highly engaged individuals we believed to be more likely to wield policy influence in the future. Many club websites were outdated, and almost no college partisans ultimately took the survey. This group made up less than 1% of our final sample, and the inclusion of this sub-sample unsurprisingly does not affect our final results (as we show in Appendix Table C4, Table C5, and Table C6).

Appendix Table A1 summarizes our sampling frame and our eventual sample of over 2,500 participants, including response rates for each sub-group. To preserve the ex-ante identification of sub-sample respondents, we created identical, parallel surveys for *each* sub-sample.

Our survey sample is clearly not representative of all political practitioners in the U.S., and those that chose to share their time with our research group almost certainly differed from those that ignored our

email. For example, the state legislators who completed our survey may have been less busy, for reasons correlated with the professionalization of their legislature, their institutional status in their chamber, or the nature of that state's legislative calendar. Alternatively, a legislator may also have been more or less likely to take our survey based on an ideological view of higher education institutions, whether or not they had college-aged children, or a specific experience with the affiliation of the authors.

While these apparent limitations in external validity are worth keeping in mind, our sample also has important strengths. Unlike convenience samples, which face persistent concerns about "bots" and falsified response information, every respondent in our survey was invited through verified, public contact information associated with an elected official or political practitioner of some sort. Moreover, as Appendix Figure A1 shows, this study is built upon a national sample with perspectives from *all 50 states and the District of Columbia*. Ultimately, we preferred that a typical respondent in our sample had real experience in a state legislature or public board of education, for example, over the opportunity to collect a representative sample of individuals that lacked these perspectives.

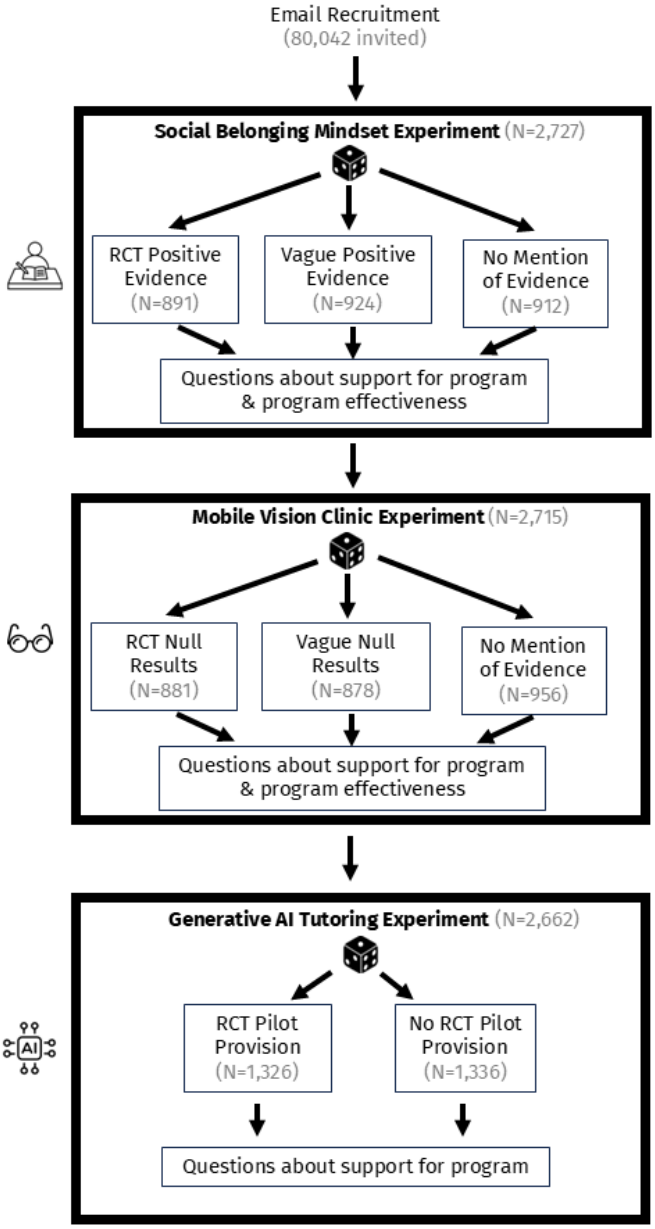
Experiments

Figure 1 provides a visual overview of our research design, which was crafted to test responses to [1] education programs that seem to work, [2] education programs with null results, and [3] education program *proposals* with an RCT pilot provision. In the first two experiments, we evaluated perceptions of program effectiveness[13] and expressions of support for adopting that program locally. In both of these experiments, respondents were randomly assigned to one of three groups. The first group received information about a program and details about evidence from RCTs, including a brief explanation of the value of random assignment. The second group also received information about the program and evidence – but in this condition the description of the evidence was left relatively vague (i.e., no mention of RCTs or any other methodology). The third group of respondents were simply informed of the program without any mention of evidence. Wherever possible, we minimized deception in our discussion of these programs; the details and findings presented to respondents in the first two experiments were from initiatives with real,

[13] In our pre-registration plan, we described the effectiveness measures as manipulation checks for brevity. Put differently, we wanted to make sure that (a) political practitioners understood evidentiary claims as a signal of greater program effectiveness and subsequently (b) shifted support in response to such evidence. Finding a shift in support *without* a change in perceived effectiveness would raise questions about whether our treatment was received as intended.

high-quality RCT evidence formatted as an accessible policy brief.[14]

Figure 1. Overview of Three Experiments



Note: The experiments were presented in a randomized order in the survey.

[14]We are not the first to randomly assign policy briefs in order to understand how best to communicate with local policymakers (e.g., Dodson et al. 2024).

In our final experiment, we introduced a new initiative designed around the preferences of undergraduate students to understand how the inclusion of an RCT pilot provision may affect support in a highly uncertain context (AI and education). To guard against unintended order effects, we also randomized which experiment each respondent viewed first, second, or third. Below we describe each experiment in turn.[15]

Experiment #1: Do Political Practitioners Respond to Successful Results?

For our first experiment, we evaluated political practitioners' responses to evidence of program success. More specifically, we introduced respondents in our survey to the positive results of the Social Belonging Mindset Program (Borman et al. 2019; Pyne and Borman 2020), an expressive writing exercise intended to counteract the academic decline that can happen as students enter middle school. Students in this program read about first-hand accounts from eighth-grade students, reflected upon feelings of social unease, and wrote about the worries that students like them might grapple with as they take tests and navigate a new learning environment. The program is inexpensive, with researcher estimates at approximately $1.35 per student.

Treatment

We used real evidence from two studies in the construction of our informational treatments (Borman et al. 2019; Pyne and Borman 2020). In choosing how we displayed research findings, we began with recommended practice in developing issue briefs for policymakers (Pennock 2023). All respondents were introduced to a description of the problem that the program sought to address, an accessible overview of how the program works, and information on where the program had been implemented. Beyond these basic details provided to the control condition, however, both the "RCT evidence" and "vague evidence" treatment groups were told that students in the program "significantly increased their GPAs" in both locations where the program had been tested. Finally, those in the "RCT evidence" group, and only that group, were also told that the program had been tested using randomized controlled trials (RCTs), along with an explanation that, "RCTs are considered the gold standard for proving a program's effectiveness" and "random assignment ensures fair comparison groups, so researchers can be confident their results aren't

[15]For each of our three experiments, we also pre-registered an explicitly exploratory question allowing respondents to share their views on the material provided in an open-ended format. We did not have any predictions associated with those open-ended responses in this study, and, while we include a few illustrative quotes in the pages to follow, we consider a more extensive, exploratory analysis of thousands of open-ended responses to be beyond the scope of this paper.

due to chance or other factors." We include the full treatment information for each of these three groups in Appendix B.

Outcome

For our primary outcome measure, we asked respondents how they would feel about a proposal to bring the social belonging mindset program to their community and presented them with a 0-to-100 scale, where lower values indicate strong opposition and higher values indicate strong support. The exact midpoint was labeled as "undecided" for respondents. In the analysis below, we consider both this continuous variable and, per our pre-analysis plan, a second, dichotomous indicator measure for responses that express support beyond the midpoint on this scale ($\geq$51).

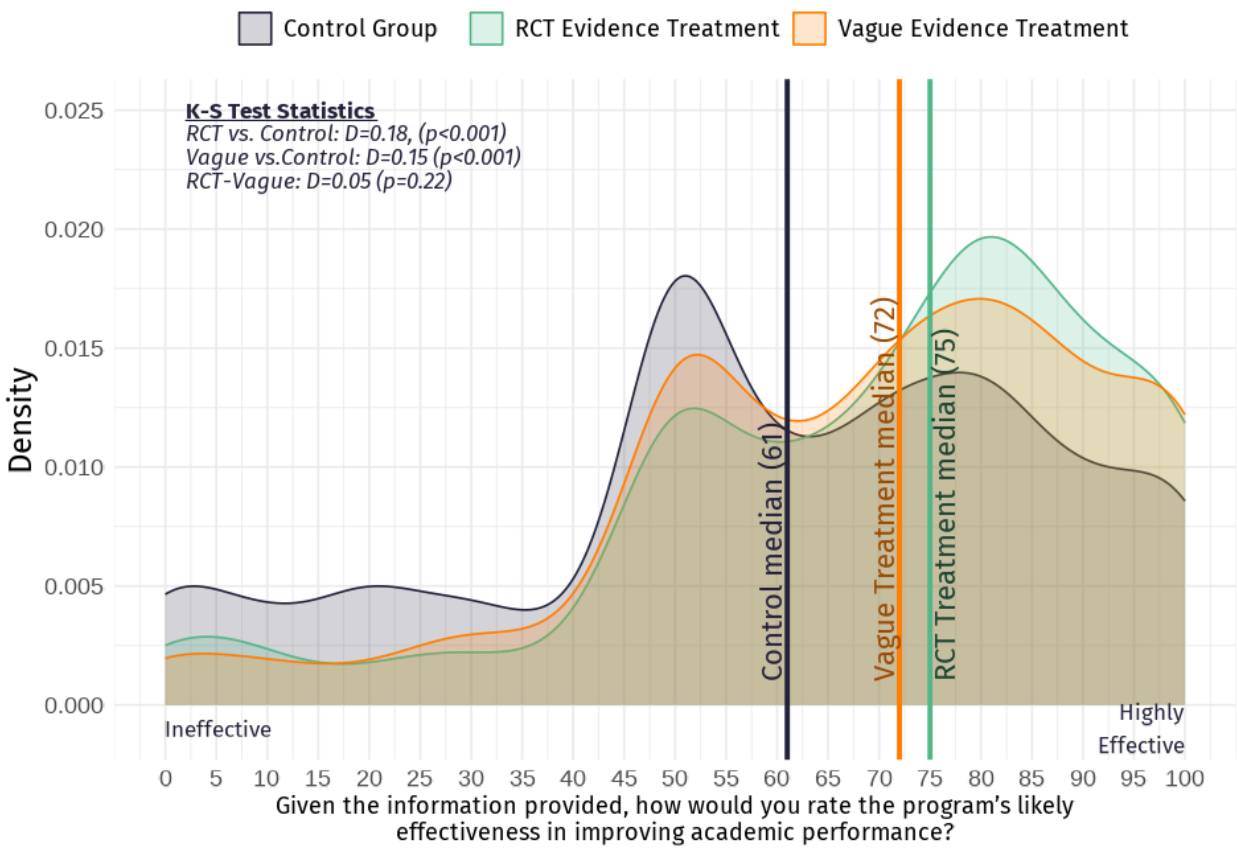
In addition to these two measures of support, we also asked respondents to evaluate the program's likely effectiveness in improving academic performance, given the information provided. If political practitioners take seriously the claims of evidence in the two treatment conditions, we anticipated a meaningful difference in perceptions of the program effectiveness. However, were we to find that the control and treatment conditions were indistinguishable from one another, this pattern would suggest that either our treatment (i.e., evidence suggesting the program succeeded) was not actually received or that political practitioners read and disregarded the validity of the research findings.[16]

Results

Figure 2, below, presents the results from our experimental analysis of perceived program effectiveness. In short, we found that political practitioners seemed to respond to the introduction of new evidence. Those in the control condition seemed cautiously optimistic about the program's impact; the median response was around 61 on a 100-point scale. By contrast, those assigned to both treatment conditions viewed the program as significantly more effective. The median perceptions of program effectiveness for both evidence treatment conditions were over ten percentage points higher (72 for "vague evidence" and 75 for "RCT-based evidence"). Kolmogorov–Smirnov (K-S) tests indicate that the overall distribution of each evidence group's responses was statistically distinct from the control condition – but the two treatment distributions (RCT vs. vague evidence) did not differ significantly from one another.

[16] Such a pattern would be consistent with evidence that politicians can heavily weigh prior attitudes even when new research is presented (Baekgaard et al. 2019).

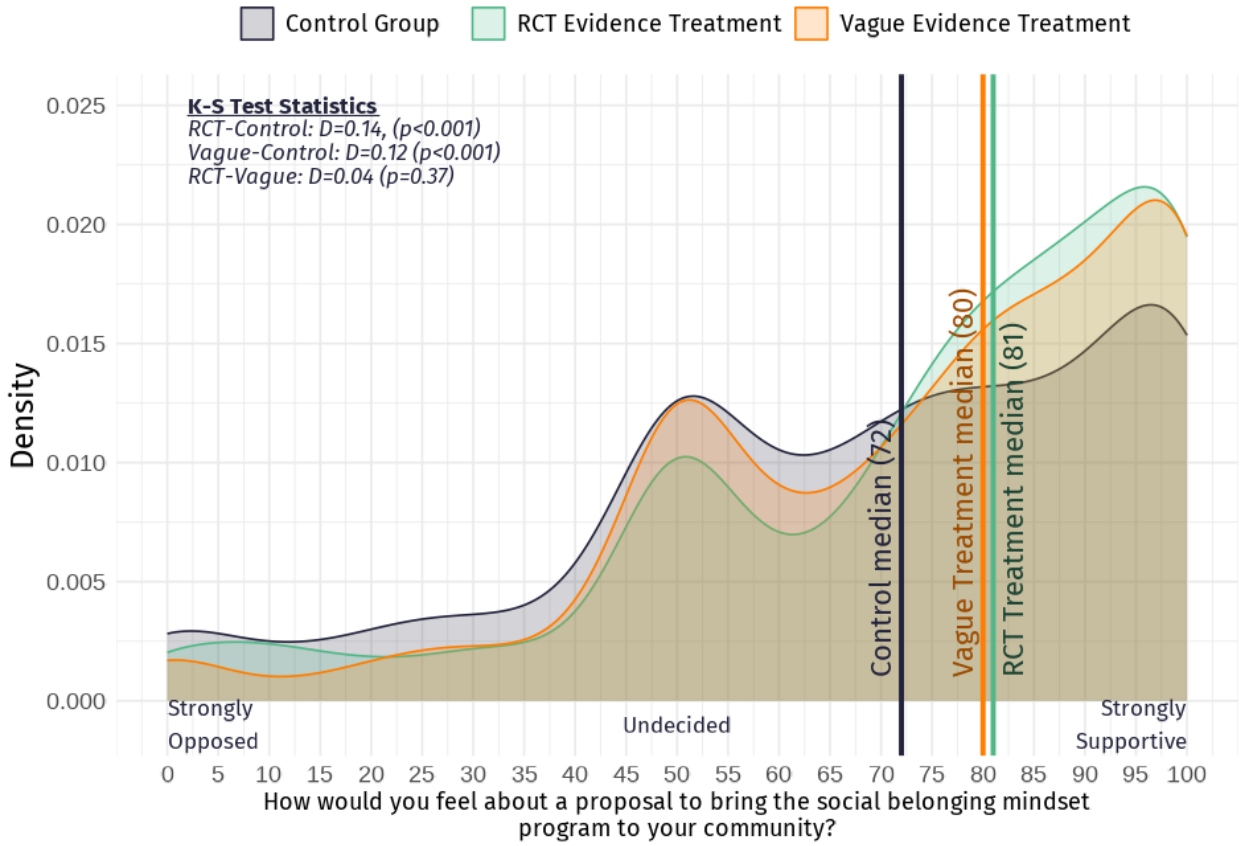
Figure 2. Effectiveness of Social Belonging Mindset Program



Note: This figure presents density plots for the full set of responses from all three experimental groups: the control group (which received no evidentiary claims), the "vague evidence" group (which received causal claims without specifying methodological details), and the "RCT evidence" group (which stated the RCT-basis for the findings, along with an accessible explanation of why random assignment is valuable). The outcome variable is the social belonging mindset program's "likely effectiveness in improving academic performance," given the information provided (0-100). K-S Test statistics are reported in the top left corner, and median reference lines are included for each of the three distributions.

Next, we found that the inclusion of causal claims affected support for education programs. Figure 3, below, presents the full distribution of our responses for each of the three treatment conditions, along with group medians and K-S statistics. The results show that evidence matters, but RCT evidence does not necessarily outperform causal claims that lack methodological details. Political practitioners expressed considerably higher levels of support (an 8–9 percentage point increase in medians) when presented with claims that the program is effective. Despite a nominally stronger expression of support among those in our "RCT evidence" group, experimental evidence did not clearly and significantly outperform vague (i.e., methodologically ambiguous) causal claims.

Figure 3. Support for Social Belonging Mindset Program



Note: This figure presents density plots for the full set of responses from all three experimental groups: the control group (which received no evidentiary claims), the "vague evidence" group (which received causal claims without specifying methodological details), and the "RCT evidence" group (which stated the RCT-basis for the findings, along with an accessible explanation of why random assignment is valuable). The outcome variable is the respondents' feelings about bringing the social belonging mindset program to their community, with values ranging from "Strongly Opposed" (0) to "Strongly Supportive" (100). K-S Test statistics are reported in the top left corner, and median reference lines are included for each of the three distributions.

We report the full set of intent-to-treat effects in Table 1 from a series of linear regression models with robust standard errors. In distributing our survey, we conducted identical, parallel surveys for each of the twelve sub-groups to preserve the *ex-ante* identification of respondent groups. The parallel administration of these identical surveys further meant that our experiment effectively used a stratified random assignment procedure, with treatment assignment occurring within each sub-sample's (identical) version of the survey. Consequently, we also include strata (i.e., sub-sample) fixed effects for all regression models

below.[17]

Table 1. Strata-FE Regression Results for Experiment #1

	Social Belonging Mindset Program (Positive Results)		
	Program Effectiveness (0–100)	Program Support (0–100)	Program Support (0,1)
Experimental Evidence Treatment	9.393***	5.497***	0.074***
	(1.213)	(1.175)	(0.020)
Vague Evidence Treatment	8.681***	5.470***	0.079***
	(1.196)	(1.140)	(0.020)
Strata Fixed Effects	✓	✓	✓
N	2,582	2,727	2,727

Note: Cell entries are OLS coefficients with robust (HC2) standard errors in parentheses. Models include strata fixed effects. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

The results in Table 1 suggest that political practitioners care about evidence, in general. Relative to the control group, both evidence treatments rated the perceived effectiveness of the program approximately nine points higher. Both treatment conditions further increased support for the educational program by 5.5 points on the 100-point scales, and the proportion of respondents indicating support at or above 51 on the 100-point scale increased by over 7 percentage points.

However, political practitioners did not significantly change their views about the program when provided with additional information indicating that the evidence was collected through a policy experiment. Wald tests comparing the RCT evidence and vague evidence coefficients in each model suggest that, across all three outcomes, the difference between the vague and RCT-evidence is not statistically significant (effectiveness: $p = 0.532$; support (0–100): $p = 0.980$; support (0,1): $p = 0.783$). We therefore cannot conclude that presenting respondents with rigorous experimental evidence changed attitudes any more than presenting the same claims without reference to RCTs. This finding holds across a continuous effectiveness

[17] Our pre-registration documents do not indicate our intention to include strata fixed effects or robust standard errors. At the time of our pre-registration, we had overlooked an element of our survey design, along with its implications for our statistical analysis. Initially, we chose to split our sample into twelve identical and parallel surveys, one for each sub-sample, in order to record ex-ante sub-sample indicators that would allow us to cleanly measure response rates. Once we realized we had effectively stratified our random assignment procedure, we committed to including strata fixed effects in all models. Given our pre-registration, we also include additional models in Appendix C that *exclude* strata (i.e., sub-sample) fixed effects.

measure, a continuous support measure, and a binary support measure. In other words, emphasizing the strong methodology used in the study did not appear to add value in communicating these important findings; clear causal claims seem to work just as well with and without reference to methodological rigor.

Experiment #2: Do Political Practitioners Respond to Null Experimental Results?

For our second experiment, we investigated the impact of high-quality null results. Here we drew upon a prominent finding from a study of mobile vision clinics in Baltimore, Maryland (Neitzel et al. 2021). In this study, mobile eye clinics visited schools, conducted free vision screenings, and performed free eye examinations for students in Baltimore City Public Schools. Students who needed glasses were given an opportunity to select their preferred frames and given free glasses (along with replacements). Surprisingly, the study led to null results on pre-registered expectations of increased academic performance; according to the researchers, "a sustained benefit was not observed over 2 years" (Neitzel et al. 2021, p. 1105). While the authors point to some suggestive findings before the end of their study, seven of the eight first-year findings fail to reach standard levels of statistical significance. After carefully reading the study's analysis, we agree with the assessment provided by Arnold Ventures, an organization that critically evaluates such studies in American public policy. This appears to be a "well-conducted RCT that produced valid findings," and the "study found no statistically significant effects on ELA or math achievement PARCC test scores" (Arnold Ventures 2023).

Treatment

Our treatment was designed in a similar way to the informational treatments described in the first experiment. All three groups received relevant context and program details, as well as some general statements suggesting that the program is important. The Baltimore City Health Commissioner was quoted as emphasizing the common-sense value of providing kids with glasses, and our infographic plainly stated that "the program helped many students receive vision care for the first time in their lives." Respondents randomly assigned to both of the treatment conditions were also informed that researchers "did <u>not</u> find statistically significant effects on state-mandated math & reading scores" after two years and that they "lacked conclusive evidence that the program had an enduring impact on student learning." Unlike the first two conditions, the information provided to our final treatment group again specified that a randomized controlled trial was used to evaluate the impact of the mobile vision clinic – and provided accompanying explanations emphasizing the importance of RCTs. (See Appendix B for the full infographics in this

experiment.)

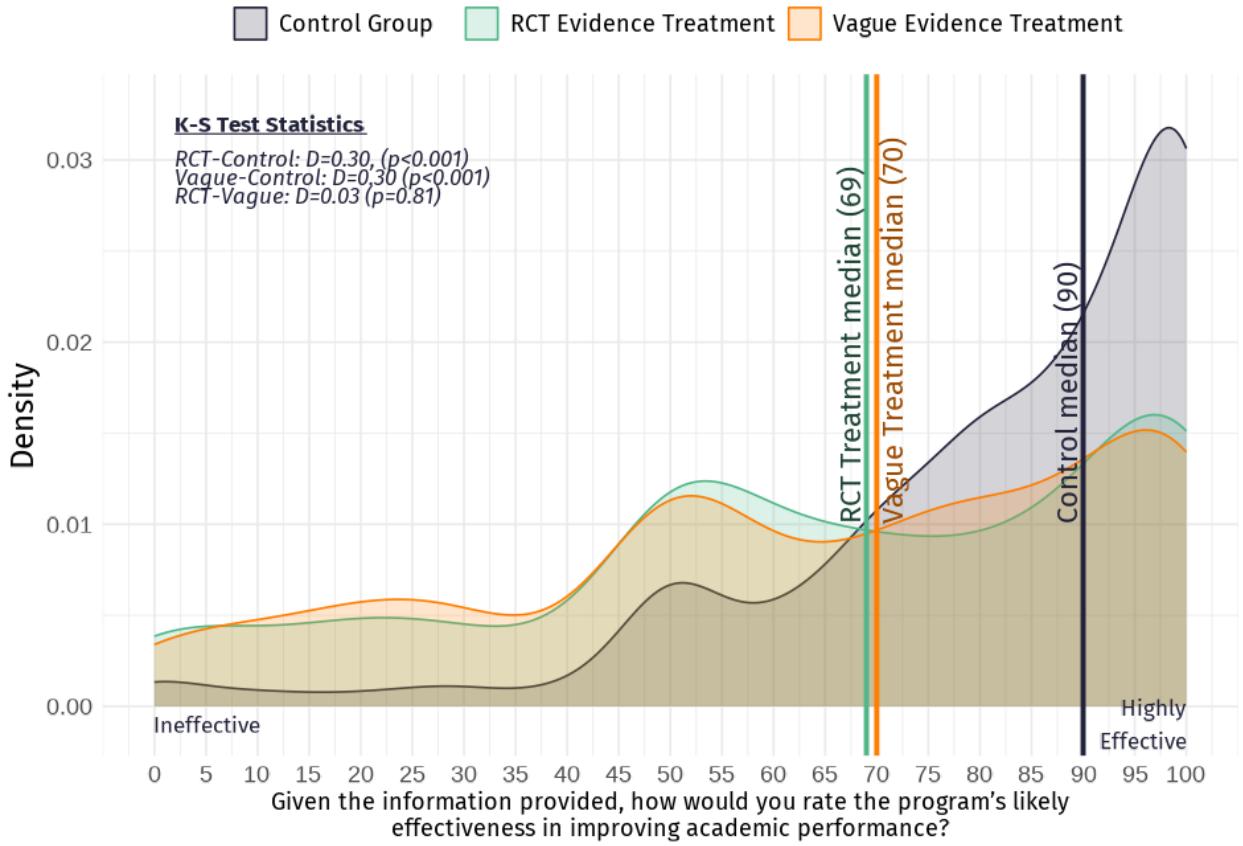
Outcomes

Our outcome measures were identical to those in the first experiment. Political practitioners recorded their reaction to this program through a continuous measure (0-100) of perceived program effectiveness, and a continuous measure (0-100) of program support. We also construct a dichotomous measure of program support using our continuous measure (1 if ≥ 51) like in our first experiment, consistent with our pre-analysis plan.

Results

Figure 4 displays the full distribution of responses to our measure of perceived program effectiveness. Policymakers viewed mobile vision clinics as less effective in both conditions that provided causal claims from prior tests. Relative to the control group, the presentation of null results lowered the median perceptions of program effectiveness by 20–21 points. Control group respondents clearly assumed that a program giving free glasses to kids would improve academic performance, but treatment group respondents reduced their estimates of the program's effectiveness after exposure to null results. Despite the significant shift in perceptions of effectiveness, the median respondent in all three conditions believed that the program was closer to highly effective than ineffective on our scale.

Figure 4. Effectiveness of Mobile Vision Clinic



While the null results presented to political practitioners seemed to shift views on the program's impact, a lack of conclusive evidence only slightly dampened their enthusiasm for the initiative. The median levels of support across all three conditions remained at or above 90 on a 100-point scale. People support giving free glasses to kids that need them – at least in the context of a survey experiment without any direct costs associated with expressions of approval. Most respondents remained supportive even after being exposed to information that the program did not provide conclusive evidence of academic benefits after two years.

Figure 5. Support for Mobile Vision Clinic

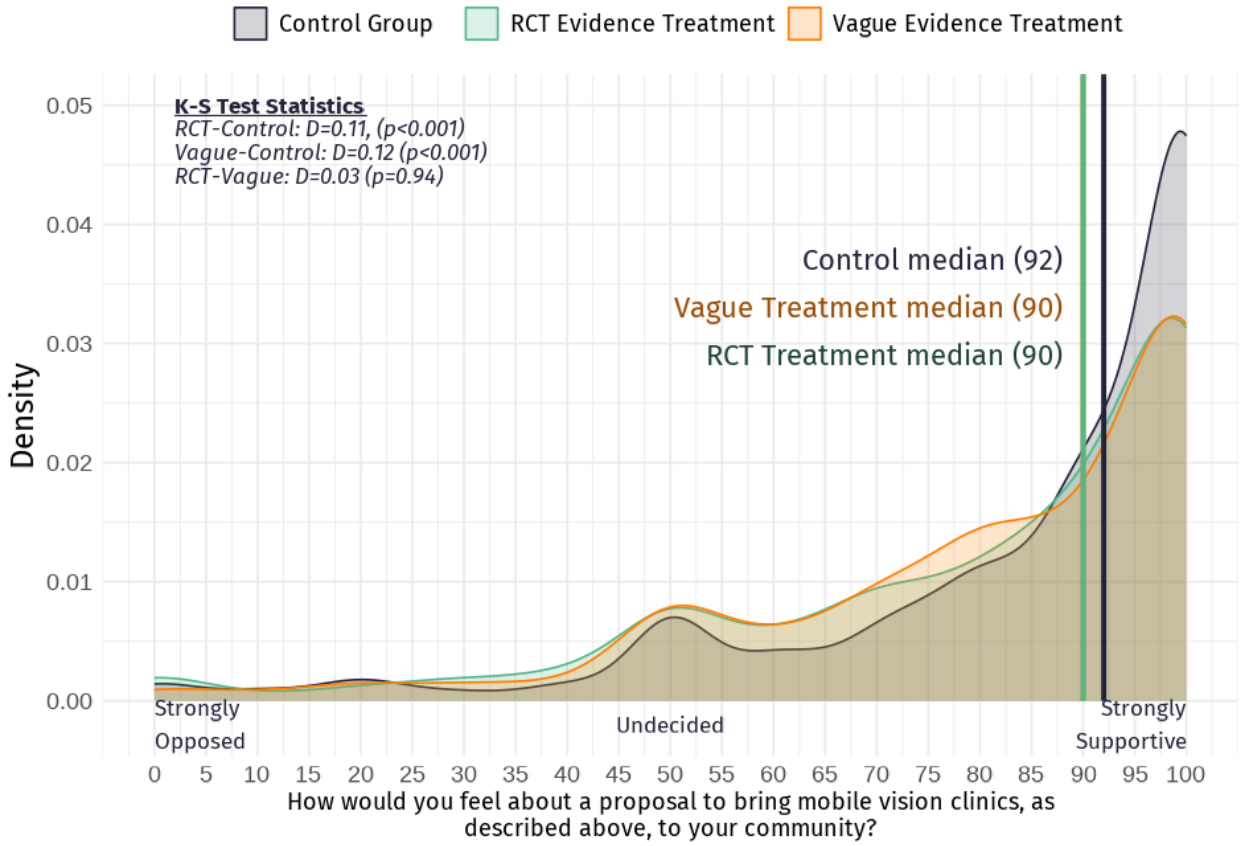


Table 2 reports our intent-to-treat effects estimated in linear regression models with strata fixed effects and robust standard errors. Practitioners, once again, clearly responded to evidence in both treatment conditions. Respondents viewed the vision clinic initiative as significantly less effective when provided with null results. Those in the RCT evidence and vague evidence groups rated the program's likely effectiveness as 17 ($p < 0.001$) and 18 ($p < 0.001$) points lower, respectively. By contrast, the effect of exposure to vague or RCT evidence on program support was smaller, with a reduction of 4–4.5 points for the two treatment conditions ($p < 0.001$). When compared to the results of our first experiment, the introduction of null results led to a *larger* shift in perceptions of program effectiveness and a much *smaller* change (-0.043 ($p < 0.01$) and -0.036 ($p < 0.05$)) in dichotomous measures of program support.[18]

[18] We note this as something to explore in future research, but it is possible that these findings reflect ceiling effects and the surprising nature of this particular study. Most importantly, we do not directly compare these attributes in a discrete choice experiment or vignette that artificially manipulates hypothetical results.

Table 2. Effects of Evidence Treatments on Perceptions and Support for the Mobile Vision Clinic

	Program Effectiveness (0–100)	Program Support (0–100)	Program Support (0,1)
Experimental Evidence Treatment	-17.152^{***}	-4.495^{***}	-0.043^{**}
	(1.243)	(1.055)	(0.016)
Vague Evidence Treatment	-18.253^{***}	-3.988^{***}	-0.036^{*}
	(1.238)	(1.012)	(0.016)
Strata Fixed Effects	✓	✓	✓
N	2,622	2,715	2,715

Note: Cell entries are OLS coefficients with robust (HC2) standard errors in parentheses. Models include strata fixed effects. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Wald tests comparing the RCT evidence and vague evidence coefficients in each model again suggest that all evidence is treated similarly in our study. Across all three outcomes, the difference between the two treatments is not statistically significant (effectiveness: $p = 0.449$; support (0–100): $p = 0.638$; support (0,1): $p = 0.654$). We therefore cannot conclude that presenting respondents with experimental evidence changed their attitudes any more than presenting the same causal claims without the indicator that those claims are backed up by a rigorous methodology. Once again, the approach to gathering evidence in the mobile vision clinic did not appear to matter as much as making a clear causal claim.

Experiment #3: Do Practitioners Respond to an Experimental Pilot Proposal?

Finally, we evaluated the presence of an RCT pilot provision in the context of a highly uncertain and salient policy domain: AI educational tools.[19] Despite the extraordinary increase in AI usage in educational contexts, our study was fielded at a time when we knew very little about how the emergence of new, AI-focused tutoring services might impact learning outcomes. In designing the experiment, undergraduate students affiliated with the authors' home institution were asked to draft a proposal on a topic of their choosing; in light of recently released AI chatbots, they chose to consider the ways in which this emergent technology might interact with tutoring services. More specifically, the student proposal focused on a popular website for creating flashcards that had recently rolled out an AI-powered tutoring service for premium accounts. In the brief, students echoed the service's emphasis on the potential for personal-

[19]Undergraduate research assistants from the authors' home institution drafted and revised the details of this proposal using the preferred terminology as of 2023: "generative AI."

ization, scalability, data-driven insights, and consistency across educational modules. This salient – and increasingly controversial – proposal seemed to students like an ideal context to understand how political practitioners value RCT pilot provisions as a means of better evaluating the uncertain consequences of a new educational tool before offering it to all students in a district.

Treatment

Our third experiment randomly assigned respondents to only two groups (compared to the three-armed designs of the first two experiments). While respondents in both the treatment and control conditions received information on the context and (from the perspective of our students) potential advantages of AI tutoring services, only the treatment condition received a pitch for an experimental (i.e., randomized) pilot program. Respondents assigned to the treatment condition were reminded of the methodological advantages of RCTs in evaluating programmatic impacts and encouraged to consider the benefits of testing the program before "committing the financial resources necessary to offer these services to all students."

Outcomes

Since this treatment included a student-drafted policy proposal, we did not solicit the respondents' thoughts on how effective the program might be. Instead, we presented them with a similar, continuous scale of program support (0-100) and constructed a dichotomous indicator variable from that data, in keeping with our pre-analysis plan.

Results

Political practitioners did not clearly change their support for the proposal when encouraged to also consider an RCT pilot program. Figure 6 provides the full distribution of responses to the proposal. Political practitioners seemed deeply uncertain about AI tutoring services at the time of our study (December 2023). The median responses for both control and treatment groups were 55 and 53, respectively – very close to the midpoint of our scale.

Figure 6. Support for AI Tutoring Services With and Without an RCT Pilot Provision

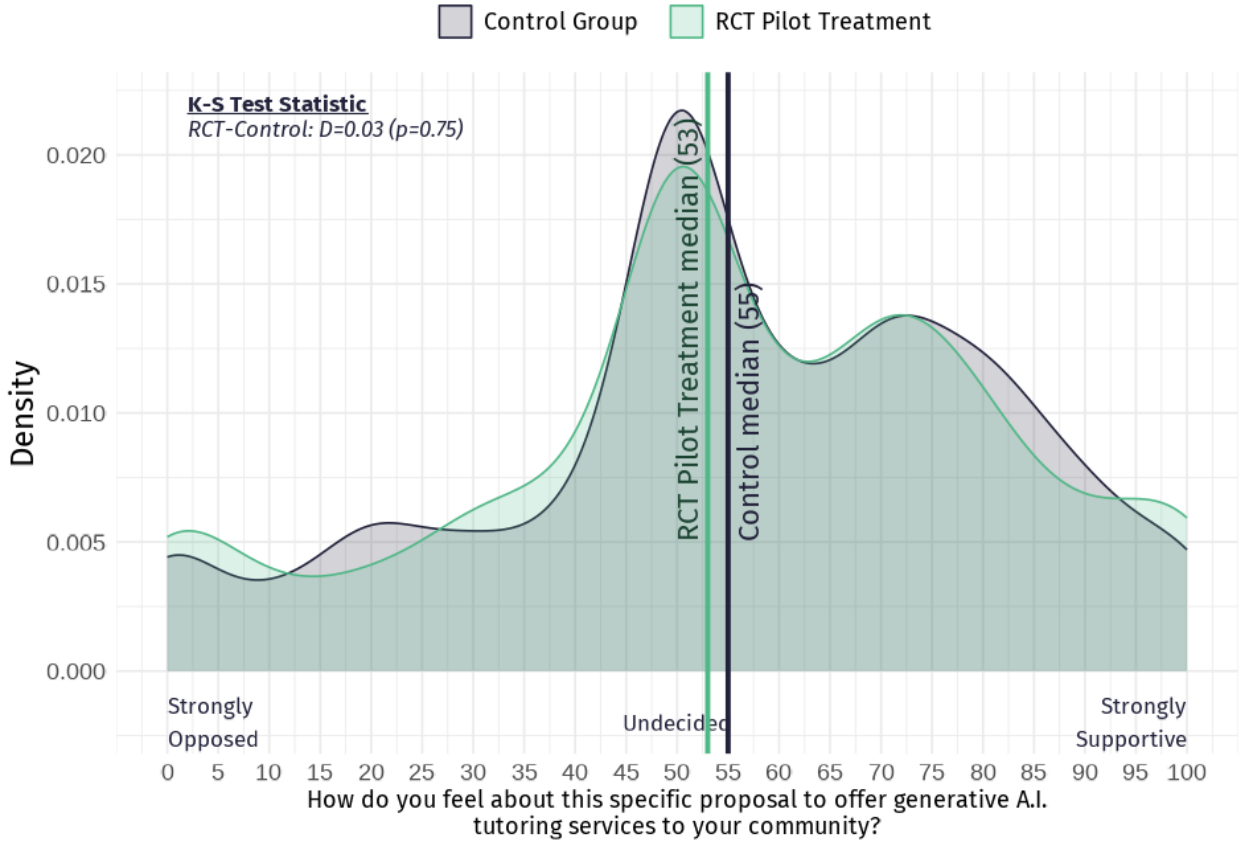


Table 3 shows the results of our linear regression models with strata fixed effects and robust standard errors. Respondents provided with language encouraging an RCT pilot provision expressed a slightly lower level of support (−0.561 for the continuous measure and −0.01 for the dichotomous measure), contrary to our expectations. However, these differences were not remotely close to statistically significant.

Table 3. Effects of RCT Pilot Treatment on Support for AI Tutoring Services

	Program Support (0–100)	Program Support (0,1)
RCT Pilot Treatment	−0.561	−0.010
	(1.000)	(0.019)
Strata Fixed Effects	✓	✓
N	2,662	2,662

Note: Cell entries are OLS coefficients with robust (HC2) standard errors in parentheses. Models include strata fixed effects. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

These findings do not provide clear evidence that an RCT pilot provision serves as a risk mitigation strategy in the minds of political practitioners. Unlike Dur et al. (2024) and Pereira et al. (2025), we do not find evidence of favorable views toward policy experimentation in this experiment. On the other hand, we also find no evidence of a penalty (i.e., "experimentation aversion"), as reported in Meyer et al. (2019) and Heck et al. (2020). Instead, political practitioners simply seemed unaffected by the proposed RCT pilot provision altogether.

Discussion

Every year, governments spend vast resources on well-intentioned and largely untested policy initiatives. We do not know if most of these programs are effective at accomplishing their goals, in part, because isolating causal effects in public policy is tremendously challenging. Public programs exist in complex social settings, and even carefully observed interventions are likely to remain thoroughly tangled up in an immeasurable number of confounding influences. Researchers have developed a range of tools to better identify program impacts, but the most reliable of these tools – randomized controlled trials – require the capacity to carefully intervene in the political world.

From a research perspective, the combined resources and coercive capacity of governments present unparalleled potential – and heightened ethical considerations – for those that wish to pursue a greater understanding of which public policies actually work. The sprawling system of American federalism, in particular, represents tens of thousands of opportunities to collaborate with governments in a position to fund, facilitate, and implement potential solutions to social problems in every corner of the country.

This research improves our understanding of the value that political practitioners place on evidence in general and policy experiments in particular. Some anecdotal accounts that predate our study suggest elected officials do not care about randomized controlled trials, *per se*. Take, for example, one state governor's perspective:

> *At a recent foundation meeting on teaching and learning, the representative of a reforming state governor spoke about a list of best educational practices his state had developed that was being disseminated to all state schools. He did not care, and he believed no governors care, about the technical quality of the research generating these lists. The main concern was to have a consensus of education researchers endorsing each practice. When asked how many of these best practices depended on results from randomized experiments, he guessed it would be zero* (Mosteller and Boruch 2002, p. 161).

The implication here seems to be that lawmakers do not care about policy experiments so much as they

care about consensus, or perhaps a veneer of authority to strengthen their preferred policy position. This approach may lead to devastating consequences. In the absence of high-quality, observable policy experiments, states may default to mere mimicry of seemingly innovative approaches (Shipan and Volden 2021), such as the mass proliferation of teaching methods that may actually *harm* young children's reading capabilities (May et al. 2024).[20] The findings presented in this study suggest that the governor quoted above is not alone in his views.

Our analysis indicates that elected officials, legislative staffers, and other political practitioners consistently respond to the presence of research-based causal claims. Americans in a position to influence education policy seem to care about evidence both when a program appears to succeed and when a program lacks conclusive evidence of success (i.e., null results). Given clear, simple, and accessible information, these individuals respond in ways that reflect the evidence put before them; our research provides evidence that political practitioners are, indeed, responsive to research.

On the other hand, we find little evidence that elected officials prefer *or penalize* policy experimentation. In our first two experiments, the random assignment of educational programs simply does not outperform vaguely specified causal claims. Stating that a program works or does not work appears to be good enough to shift preferences – even if those claims are not clearly backed by a careful causal identification strategy. The introduction of an RCT pilot proposal in our final experiment fails to clearly move support, but political practitioners do not clearly penalize proposals that randomly assign AI tutoring services for children – an increasingly controversial proposal, given the uncertain consequences of this emergent technology. Instead of experimental aversion (Meyer et al. 2019; Heck et al. 2020) or enthusiasm (Dur et al. 2024), our study suggests experimental indifference among political practitioners.

This set of findings calls for new research into the politics of evidence. Our study presents information in a controlled survey environment, rather than the informationally crowded environments in which administrators are asked to make decisions. As others have noted, opposition to policy experiments may be highly conditional, and new research may advance our understanding of these views by more directly measuring openness to experimentation as an outcome of interest (Cardon 2022). Additional descriptive work, following (Cardon and Lopoo 2024), can help us to better understand varying degrees of openness

[20]This case was popularized by the podcast *Sold a Story*; within three years of that podcast's reporting, 26 states passed legislation abandoning "three-cueing" reading methods (Peak 2025). While a case study is beyond the scope of this paper, this pattern of abandonment may be consistent with the theoretical predictions in Volden, Ting, and Carpenter (2008) and the empirical findings in (Volden 2016) .

to policy experiments across distinct populations of politically influential Americans.

Future work might also build upon recent efforts to better understand belief polarization (Lee et al. 2021) and the specific forms of decision aids that may be useful in communicating evidence to policymakers (Toma and Bell 2024). The informational treatments in our studies were built upon one set of recommended practices in crafting a particular form of written communication to lawmakers, but evidence is frequently communicated to policymakers through presentations, hearings, professional networks, and many other modes of communication not captured in our experimental design. Given our findings, we are particularly interested in understanding how clean case studies, such as those stories of success presented in Hart and Yohannes (2019), might help begin conversations with those in a position to collaborate on policy experiments. Each of these presents meaningful opportunities to advance our understanding of scientific communication in the halls of political power.

More work is also needed to better map the laboratories of American democracy. Despite the extraordinary rise of policy experimentation, we lack a comprehensive picture of *where* randomized trials in public policy take place. The capacity to conduct rigorous field experiments is not unlikely to be evenly distributed throughout the United States, which is important in light of prior evidence that policymakers seem to care quite a lot about external validity considerations (Crawfurd et al. 2025). If, for example, policy experimentation disproportionately occurs in metropolitan areas, good policies may not diffuse through rural regions wary of adopting a program from a dissimilar context. Careful descriptive data on government-supported trials can shed light on *which* political communities stand to benefit from such interventions – and which communities carry the risks of testing unproven initiatives.

The proliferation of applied research units embedded within local governments (e.g., the Lab @ DC) further presents an opportunity to better understand how the lessons of clearly observable policy experiments diffuse across *and within* state boundaries. Importantly, this work should not stop with an understanding of program adoption; researchers should continue to consider the ways in which policy experiments are framed and understood in public forms of political communication (Gilardi et al. 2021). Given recent findings that ideological biases can be mitigated by emphasizing the success of policies and the presence of co-partisan experimentation (Butler et al. 2017), this line of inquiry seems particularly important in an era of deep partisan conflict.

References

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30. https://doi.org/10.1257/jep.24.2.3.

Arnold Ventures. 2023. "A Randomized Controlled Trial of the Vision for Baltimore (V4B) Project, Which Provides Free Eyeglasses to Disadvantaged Students in Elementary and Middle School." https://www.arnoldventures.org/stories/a-randomized-controlled-trial-of-the-vision-for-baltimore-v4b-project-which-provides-free-eyeglasses-to-disadvantaged-students-in-elementary-and-middle-school.

Associated Press. 2021. "TV Academy Revokes Cuomo's Emmy amid Scandal, Resignation." *AP News*, August 24, 2021. https://apnews.com/article/andrew-cuomo-entertainment-science-coronavirus-pandemic-emmy-awards-4e2430858d492b62b877ffeb2e10a83c.

Baron, Jon. 2018. "A Brief History of Evidence-Based Policy." *The Annals of the American Academy of Political and Social Science* 678: 40–50.

Baekgaard, Martin, Julian Christensen, Casper Mondrup Dahlmann, Asbjørn Mathiasen, and Niels Bjørn Grund Petersen. 2019. "The Role of Evidence in Politics: Motivated Reasoning and Persuasion among Politicians." *British Journal of Political Science* 49(3): 1117–40. https://doi.org/10.1017/S0007123417000084.

Benartzi, Shlomo, John Beshears, Katherine L. Milkman, Cass R. Sunstein, Richard H. Thaler, Maya Shankar, Will Tucker-Ray, William J. Congdon, and Steven Galing. 2017. "Should Governments Invest More in Nudging?" *Psychological Science* 28 (8): 1041–55. https://doi.org/10.1177/0956797617702501.

Bergan, Daniel E., Kelechi Amakoh, Dustin Carnahan, Dan Hiaeshutter-Rice, and Lu Teinowitz. 2026. "Are Policymakers Persuaded in Parallel?: Policymakers, Evidence, and Motivated Reasoning." *American Politics Research* 54(1): 80–90. https://doi.org/10.1177/1532673X251367389.

Borman, Geoffrey D., Christopher S. Rozek, Jaymes Pyne, and Paul Hanselman. 2019. "Reappraising Academic and Social Adversity Improves Middle School Students' Academic Achievement, Behavior, and Well-Being." *Proceedings of the National Academy of Sciences* 116 (33): 16286–91. https://doi.org/10.1073/pnas.1820317116.

Briscese, Guglielmo, and John List. 2024. "Toward an Understanding of the Political Economy of Using Field Experiments in Policymaking." NBER Working Paper 33239. Cambridge, MA: National Bureau of Economic Research. http://www.nber.org/papers/w33239.

Butler, Daniel M., Craig Volden, Adam M. Dynes, and Boris Shor. 2017. "Ideology, Learning, and Policy Diffusion: Experimental Evidence." *American Journal of Political Science* 61(1): 37–49. https://doi.org/10.1111/ajps.12213.

Cairney, Paul. 2017. "The Politics of Evidence-Based Policy Making." In *Oxford Research Encyclopedia of Politics*, ed. Erin Hannah. Oxford University Press. New York, NY. https://doi.org/10.1093/acrefore/9780190228637.013.268.

Cairney, Paul, and Kathryn Oliver. 2017. "Evidence-Based Policymaking Is Not like Evidence-Based Medicine, so How Far Should You Go to Bridge the Divide between Evidence and Policy?" *Health Research Policy and Systems* 15(1): 35. https://doi.org/10.1186/s12961-017-0192-x.

Cardon, Emily, and Leonard Lopoo. 2024. "Randomized Controlled Trial Aversion among Public Sector Leadership: A Survey Experiment." *Evaluation Review* 48(4): 579–609. https://doi.org/10.1177/0193841X231193483.

Cardon, Emily. 2022. "Public Administrator Aversion to Randomized Controlled Trials." PhD diss., Syracuse University. https://surface.syr.edu/record/7159?v=pdf.

Cartwright, Nancy, and Eileen Munro. 2010. "The Limitations of Randomized Controlled Trials in Predicting Effectiveness." *Journal of Evaluation in Clinical Practice* 16 (2): 260–66.

Connolly, Paul, Ciara Keenan, and Karolina Urbanska. 2018. "The Trials of Evidence-Based Practice in Education: A Systematic Review of Randomised Controlled Trials in Education Research 1980–2016." *Educational Research* 60(3): 276–91. https://doi.org/10.1080/00131881.2018.1493353.

Chlond, Bettina, Timo Goeschl, and Johannes Lohse. 2025. "Revealed Preferences for Policy Experiments." SSRN Working Paper. https://dx.doi.org/10.2139/ssrn.5235601.

Crawfurd, Lee, Susannah Hares, Ana Minardi, and Justin Sandefur. 2025. "Understanding Education Policy Preferences: Survey Experiments with Policymakers in 35 Developing Countries." *World Development* 196: 107140. https://doi.org/10.1016/j.worlddev.2025.107140.

DellaVigna, Stefano, Woojin Kim, and Elizabeth Linos. 2024. "Bottlenecks for Evidence Adoption." *Journal of Political Economy* 132 (8): 2748–89. https://doi.org/10.1086/729447.

Dodson, Elizabeth A., Renee G. Parks, Rebekah R. Jacob, Ruopeng An, Amy A. Eyler, Nathan Lee, Alexandra B. Morshed, et al. 2024. "Effectively Communicating with Local Policymakers: A Randomized Trial of Policy Brief Dissemination to Address Obesity." *Frontiers in Public Health* 12: 1246897. https://doi.org/10.3389/fpubh.2024.1246897.

Dur, Robert, Arjan Non, Paul Prottung, and Benedetta Ricci. 2024. "Who's Afraid of Policy Experiments?" *The Economic Journal* 135(666): 538–55. https://doi.org/10.1093/ej/ueae090.

Fagan, E. J., and Zachary A. McGee. 2022. "Problem Solving and the Demand for Expert Information in Congress." *Legislative Studies Quarterly* 47 (1): 53–77. https://doi.org/10.1111/lsq.12323.

Geller, Martinne, and Jonathan Stempel. 2020. "Disinfectant Makers Steer Consumers Away from Trump's Coronavirus Comments." *Reuters*, April 24, 2020. https://www.reuters.com/article/idUSKCN2261DD/.

Gilardi, Fabrizio, Charles R. Shipan, and Bruno Wüest. 2021. "Policy Diffusion: The Issue-Definition Stage." *American Journal of Political Science* 65(1): 21–35. https://doi.org/10.1111/ajps.12521.

Gilardi, Fabrizio, and Fabio Wasserfallen. 2019. "The Politics of Policy Diffusion." *European Journal of Political Research* 58 (4): 1245–56. https://doi.org/10.1111/1475-6765.12326.

Greenberg, David, and Mark Shroder. 1997. *The Digest of Social Experiments*. 2nd ed. Washington, DC: Urban Institute Press.

Gueron, Judith M. 2002. "The Politics of Random Assignment: Implementing Studies and Affecting Policy." In *Evidence Matters: Randomized Trials in Education Research*, eds. Frederick Mosteller and Robert Boruch. Washington, DC: Brookings Institution Press, 15–49. https://muse.jhu.edu/book/63815.

Gueron, Judith M., and Howard Rolston. 2013. *Fighting for Reliable Evidence*. New York: Russell Sage Foundation. https://www.jstor.org/stable/10.7758/9781610448130.

Hart, Nick, and Meron Yohannes, eds. 2019. *Evidence Works: Cases Where Evidence Meaningfully Informed Policy*. Washington, DC: Bipartisan Policy Center. https://bipartisanpolicy.org/wp-content/uploads/2019/06/Evidence-Works-Cases-Where-Evidence-Meaningfully-Informed-Policy.pdf.

Haskins, Ron, and Greg Margolis. 2014. *Show Me the Evidence: Obama's Fight for Rigor and Results in Social Policy*. Washington, DC: Brookings Institution Press.

Heck, Patrick R., Christopher F. Chabris, Duncan J. Watts, and Michelle N. Meyer. 2020. "Objecting to Experiments Even While Approving of the Policies or Treatments They Compare." *Proceedings of the National Academy of Sciences* 117(32): 18948–50. https://doi.org/10.1073/pnas.2009030117.

Hedges, Larry V., and Jacob Schauer. 2018. "Randomised Trials in Education in the USA." *Educational Research* 60 (3): 265–75. https://doi.org/10.1080/00131881.2018.1493350.

Hjort, Jonas, Diana Moreira, Gautam Rao, and Juan Francisco Santini. 2021. "How Research Affects Policy: Experimental Evidence from 2,150 Brazilian Municipalities." *American Economic Review* 111(5): 1442–80. doi:10.1257/aer.20190830.

International Academy of Television Arts and Sciences. 2020. "New York Governor Andrew M. Cuomo to Receive 2020 International Emmy Founders Award." November 20, 2020. https://web.archive.org/web/20201120230834/https://www.iemmys.tv/new-york-governor-andrew-m-cuomo-to-receive-2020-international-emmy-founders-award/.

Jatteau, Arthur. 2018. "The Success of Randomized Controlled Trials: A Sociographical Study of the Rise of J-PAL to Scientific Excellence and Influence." *Historical Social Research / Historische Sozialforschung* 43 (3 (165)): 94–119.

Knox, Virginia, Carolyn J. Hill, and Gordon Berlin. 2018. "Can Evidence-Based Policy Ameliorate the Nation's Social Problems?" *The Annals of the American Academy of Political and Social Science* 678: 166–79.

Lee, Nathan, Brendan Nyhan, Jason Reifler, and D.J. Flynn. 2021. "More Accurate, But No Less Polarized: Comparing the Factual Beliefs of Government Officials and the Public." *British Journal of Political Science* 51(3): 1315–22. https://doi.org/10.1017/S000712342000037X.

Lee, Nathan. 2022. "Do Policy Makers Listen to Experts? Evidence from a National Survey of Local and State Policy Makers." *American Political Science Review* 116(2): 677–88. https://doi.org/10.1017/S0003055421000800.

May, Henry, Aly Blakeney, Pragya Shrestha, Mia Mazal, and Nicole Kennedy. 2024. "Long-Term Impacts of Reading Recovery through 3rd and 4th Grade: A Regression Discontinuity Study." *Journal of Research on Educational Effectiveness* 17(3): 433–58. https://doi.org/10.1080/19345747.2023.2209092.

Masset, Edoardo, Marie Gaarder, Penelope Beynon, and Christelle Chapoy. 2013. "What Is the Impact of a Policy Brief? Results of an Experiment in Research Dissemination." *Journal of Development Effectiveness* 5(1): 50–63. https://doi.org/10.1080/19439342.2012.759257.

Matthews, Abigail A., and Frederick J. Boehmke. 2026. "Researchers in the Laboratories of Democracy: A Quarter Century of Innovations in Studying Policy Diffusion in State Politics & Policy Quarterly." *State Politics & Policy Quarterly* 26(2): 185–205. https://doi.org/10.1017/spq.2026.10020.

Mehmood, Sultan, Shaheen Naseer, and Daniel L. Chen. 2024a. "Training Policymakers in Econometrics." https://users.nber.org/~dlchen/papers/Training_Policy_Makers_in_Econometrics.pdf.

Mehmood, Sultan, Shaheen Naseer, and Daniel L. Chen. 2024b. "Training Policymakers in Econometrics II: Tax Officers and Fiscal State Capacity." https://users.nber.org/~dlchen/papers/Training_Policy_Makers_in_Econometrics_II.pdf.

Meyer, Michelle N., Patrick R. Heck, Geoffrey S. Holtzman, Stephen M. Anderson, William Cai, Duncan J. Watts, and Christopher F. Chabris. 2019. "Objecting to Experiments That Compare Two Unobjectionable Policies or Treatments." *Proceedings of the National Academy of Sciences* 116(22): 10723–28. https://doi.org/10.1073/pnas.1820701116.

Mosteller, Frederick F., and Robert F. Boruch. 2002. *Evidence Matters: Randomized Trials in Education Research*. Washington, DC: Brookings Institution Press.

Nakajima, Nozomi. 2021. "Evidence-Based Decisions and Education Policymakers." https://nozominakajima.github.io/files/nakajima_policymaker.pdf.

Neitzel, Amanda J., Betsy Wolf, Xinxing Guo, Ahmed F. Shakarchi, Nancy A. Madden, Michael X. Repka, David S. Friedman, and Megan E. Collins. 2021. "Effect of a Randomized Interventional School-Based Vision Program on Academic Performance of Students in Grades 3 to 7." *JAMA Ophthalmology* 139 (10): 1104–14. https://doi.org/10.1001/jamaophthalmol.2021.3544.

Nobel Prize Outreach. 2019. "The Prize in Economic Sciences 2019." NobelPrize.org. https://www.nobelprize.org/prizes/economic-sciences/2019/press-release/.

Office of Evaluation Sciences (OES). 2021. "Understanding and Improving How Policymakers Respond to Program Impact." Washington, DC: U.S. General Services Administration. https://oes.gsa.gov/assets/abstracts/2101-abstract-gsa-policymakers.pdf.

Office of Evaluation Sciences (OES). 2025. "Evaluating Generative Artificial Intelligence (GenAI) Chat Tools: Learning About U.S. General Services Administration Employees' GenAI Usage and Support Needs." Washington, DC: U.S. General Services Administration. https://oes.gsa.gov/assets/abstracts/evaluating-genai-chat-tools.pdf.

Peak, Christopher. 2025. "New Reading Laws Sweep the Nation Following *Sold a Story*." *APM Reports*, October 16, 2025. https://www.apmreports.org/story/2025/10/16/legislators-reading-laws-sold-a-story.

Pennock, Andrew. 2023. *The CQ Press Writing Guide for Public Policy*. 2nd ed. Washington, DC: CQ Press.

Penuel, William R., Derek C. Briggs, Kristen L. Davidson, Corinne Herlihy, David Sherer, Heather C. Hill, Caitlin Farrell, and Anna-Ruth Allen. 2017. "How School and District Leaders Access, Perceive, and Use Research." *AERA Open* 3(2). https://doi.org/10.1177/2332858417705370.

Pereira, Miguel M., Martin Baekgaard, Nathalie Giger, and Lior Sheffer. 2025. "Politicians and Evidence." https://osf.io/preprints/osf/q7h2c_v2.

Pyne, Jaymes, and Geoffrey Borman. 2020. "Replicating a Scalable Intervention That Helps Students Reappraise Academic and Social Adversity During the Transition to Middle School." *Journal of Research on Educational Effectiveness* 13 (4): 652–78. https://doi.org/10.1080/19345747.2020.1784330.

Shea, Robert. 2004. "What Constitutes Strong Evidence of a Program's Effectiveness?" Washington, DC: Office of Management and Budget.

Shipan, Charles R., and Craig Volden. 2021. *Why Bad Policies Spread (and Good Ones Don't)*. New York: Cambridge University Press.

Slavin, Robert E. 2002. "Evidence-Based Education Policies: Transforming Educational Practice and Research." *Educational Researcher* 31 (7): 15–21. https://doi.org/10.3102/0013189X031007015.

Stevenson, Megan T. 2023. "Cause, Effect, and the Structure of the Social World." https://ssrn.com/abstract=4445710.

Toma, Mattie, and Elizabeth Bell. 2024. "Understanding and Increasing Policymakers' Sensitivity to Program Impact." *Journal of Public Economics* 234: 105096. https://doi.org/10.1016/j.jpubeco.2024.105096.

Vivalt, Eva, and Aidan Coville. 2023. "How Do Policymakers Update Their Beliefs?" *Journal of Development Economics* 165: 103121. https://doi.org/10.1016/j.jdeveco.2023.103121.

Vivalt, Eva, Aidan Coville, and Sampada Kc. 2025. "Local Knowledge, Formal Evidence, and Policy Decisions." *Journal of Development Economics* 174: 103425. https://doi.org/10.1016/j.jdeveco.2024.103425.

Volden, Craig. 2006. "States as Policy Laboratories: Emulating Success in the Children's Health Insurance Program." *American Journal of Political Science* 50(2): 294–312. https://doi.org/10.1111/j.1540-5907.2006.00185.x.

Volden, Craig. 2016. "Failures: Diffusion, Learning, and Policy Abandonment." *State Politics & Policy Quarterly* 16(1): 44–77. https://doi.org/10.1177/1532440015588910.

Volden, Craig, Michael M. Ting, and Daniel P. Carpenter. 2008. "A Formal Model of Learning and Policy Diffusion." *American Political Science Review* 102(3): 319–32. https://doi.org/10.1017/S0003055408080271.

Zelizer, Adam. 2018. "How Responsive Are Legislators to Policy Information? Evidence from a Field Experiment in a State Legislature." *Legislative Studies Quarterly* 43(4): 595–618. https://doi.org/10.1111/lsq.12206.

Supplemental Materials for

"Do Political Practitioners Care About Policy Experiments?"

Contents

A Sample **S2**

B Survey Instrument Screenshots **S7**

C Additional Analyses **S15**

*Andrew Clarke (clarkeaj@lafayette.edu) is the corresponding author. This research was approved by the Institutional Review Board of Lafayette College on December 4, 2023 (Proposal Number: AY2324-37). Our pre-analysis plan (AsPredicted #155752) was registered just before our recruitment email was distributed, on December 18, 2023. In preparing this manuscript, the authors occasionally used AI tools (Claude and Overleaf AI) for copyediting of the text and the proofreading and tidying of hand-drafted R scripts; Appendix C was an outlier to this rule, as we relied more heavily on Claude-generated first-draft code for some robustness checks (e.g., randomization inference). This research would not have been possible without the tremendous research assistance of the students in the Lafayette College Gov Lab and funding from the Landis Center for Community Engagement. We are also indebted to Stephanie Chan, Dylan Groves, Adam Biener, and the Lafayette College Department of Economics for helpful feedback and guidance at early and formative stages in this project.

A Sample

Table A1. Summary of Respondents Completing *All Three* Experiments, by Sub-Sample

	Invited	Respondents	Response Rate	% of Overall Sample
School Board Members	53,518	1,548	2.9%	60.1%
State Lawmakers (Democrats)	3,122	99	3.2%	3.8%
State Lawmakers (Republicans)	3,822	67	1.8%	2.6%
State Lawmakers (Others)	68	1	1.5%	0.04%
State Legislative Staffers (Democrats)	4,162	215	5.2%	8.3%
State Legislative Staffers (Republicans)	3,260	105	3.2%	4.1%
State Legislative Staffers (Others)	295	11	3.7%	0.4%
Local Party Leaders (Democrats)	4,978	291	5.8%	11.3%
Local Party Leaders (Republicans)	3,619	134	3.7%	5.2%
Political Candidates (NC)	2,205	89	4.0%	3.5%
College Democrats	511	7	1.4%	0.3%
College Republicans	482	8	1.7%	0.3%
Full Sample	80,042	2,575	3.2%	100%

Notes: This table summarizes respondents that completed all outcome measures for all three experiments. Invited counts are the number of individuals invited after removing bounced and broken email addresses and duplicates across sub-samples. The response rate is respondents divided by invited. The percent of the full sample is respondents divided by 2,575, the total number of unique respondents that completed *all three* experiments.

Table A2. Summary of Respondents Completing *At Least One* Experiment, by Sub-Sample

	Invited	Respondents	Response Rate	% of Overall Sample
School Board Members	53,518	1,686	3.2%	60.0%
State Lawmakers (Democrats)	3,122	111	3.6%	4.0%
State Lawmakers (Republicans)	3,822	76	2.0%	2.7%
State Lawmakers (Others)	68	2	2.9%	0.1%
State Legislative Staffers (Democrats)	4,162	239	5.7%	8.5%
State Legislative Staffers (Republicans)	3,260	115	3.5%	4.1%
State Legislative Staffers (Others)	295	12	4.1%	0.4%
Local Party Leaders (Democrats)	4,978	309	6.2%	11.0%
Local Party Leaders (Republicans)	3,619	149	4.1%	5.3%
Political Candidates (NC)	2,205	94	4.3%	3.3%
College Democrats	511	8	1.6%	0.3%
College Republicans	482	8	1.7%	0.3%
Full Sample	80,042	2,809	3.5%	100%

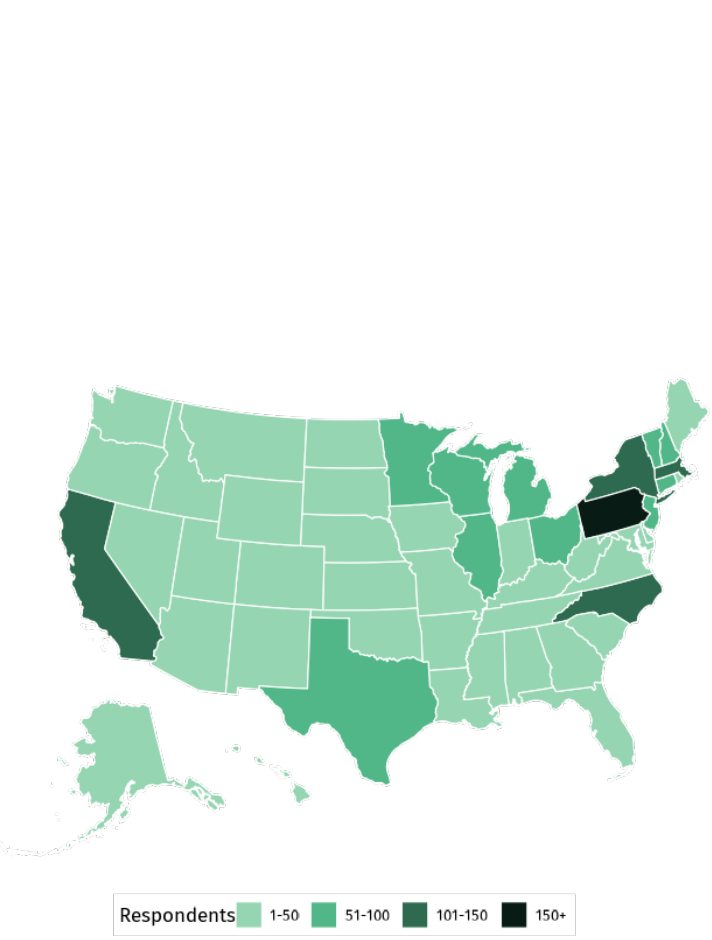
Notes: This table summarizes respondents that completed both outcome measures in at least one of our three experiments (the effectiveness and support questions in the social belonging and mobile vision clinics experiments, and the support question in the generative AI tutoring experiment). Invited counts are the number of individuals invited after removing bounced and broken email addresses and duplicates across sub-samples, and are identical to those in Table A1. The response rate is respondents divided by invited. The percent of the full sample is respondents divided by 2,809, the total number of unique respondents.

Table A3. Comparison of Our Sample to Lee (2022)'s sample in *APSR*

	Lee (2022)	Our Sample
Total respondents	690	2,809
Respondents per experiment (or issue)	456–547	2,582–2,662
Invited	Not reported	80,042
Response rate	Not reported	3.5%
States represented	50	50 + DC
Respondents by type		
Elected municipal officials	252	1,686
State legislators	90	189
State legislative staffers	190	366
Elected county officials	158	–
County party leaders	–	458
Political candidates (NC)	–	94
College partisans	–	16

Notes: Comparison to Nathan Lee. 2022. "Do Policy Makers Listen to Experts? Evidence from a National Survey of Local and State Policy Makers," *American Political Science Review* 116(2): 677–688. The Lee (2022) values are taken from the article and its online appendix. Note that Lee (2022) invited a random sample of local and state policymakers drawn from CivicPulse's contact list of elected and appointed officials in all township, municipality, county, and state legislative districts with populations of 1,000 or more. We were unable to locate the number of invited policymakers, which precluded our ability to calculate a response rate. All sample counts from our study reflect the number of respondents that completed both outcome measures in at least one experiment (see Table A2). Respondents per experiment (or issue) indicate the number of respondents answering the outcome questions in each of our three experiments and in each of Lee's three policy issues. Lee randomly assigned each respondent to two of his three issues (his online appendix reports 476, 456, and 547 respondents for needle exchanges, GMO bans, and rent control, respectively), whereas our respondents were randomized in all three of our experiments. Elected local officials are municipal officials in Lee (2022) and elected members of public boards of education in our sample.

Figure A1. Geographic Distribution of Sample



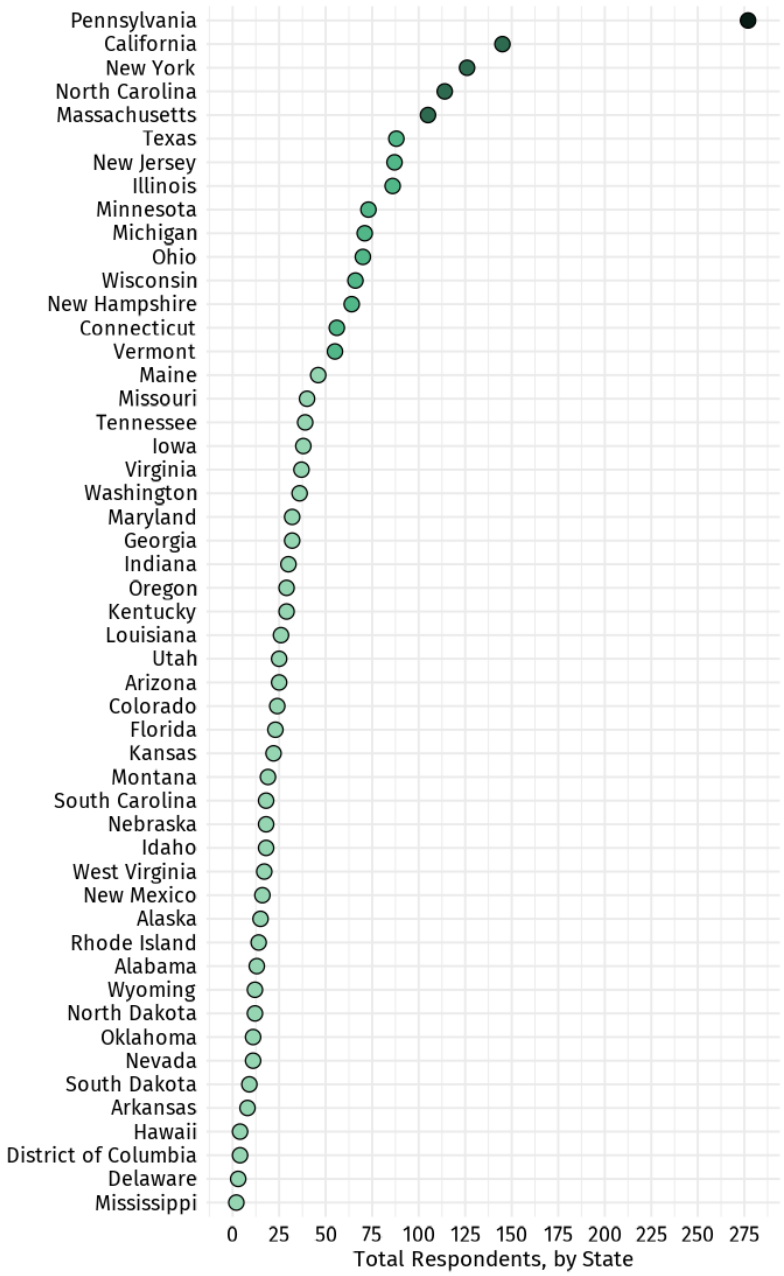


Table A4: Survey Respondents by State

State	Respondents	State	Respondents
Pennsylvania	277	Louisiana	26
California	145	Arizona	25
New York	126	Utah	25
North Carolina	114	Colorado	24
Massachusetts	105	Florida	23
Texas	88	Kansas	22
New Jersey	87	Montana	19
Illinois	86	Idaho	18
Minnesota	73	Nebraska	18
Michigan	71	South Carolina	18
Ohio	70	West Virginia	17
Wisconsin	66	New Mexico	16
New Hampshire	64	Alaska	15
Connecticut	56	Rhode Island	14
Vermont	55	Alabama	13
Maine	46	North Dakota	12
Missouri	40	Wyoming	12
Tennessee	39	Nevada	11
Iowa	38	Oklahoma	11
Virginia	37	South Dakota	9
Washington	36	Arkansas	8
Georgia	32	District of Columbia	4
Maryland	32	Hawaii	4
Indiana	30	Delaware	3
Kentucky	29	Mississippi	2
Oregon	29		

B Survey Instrument Screenshots

Figure B1. Experiment #1 Control Infographic

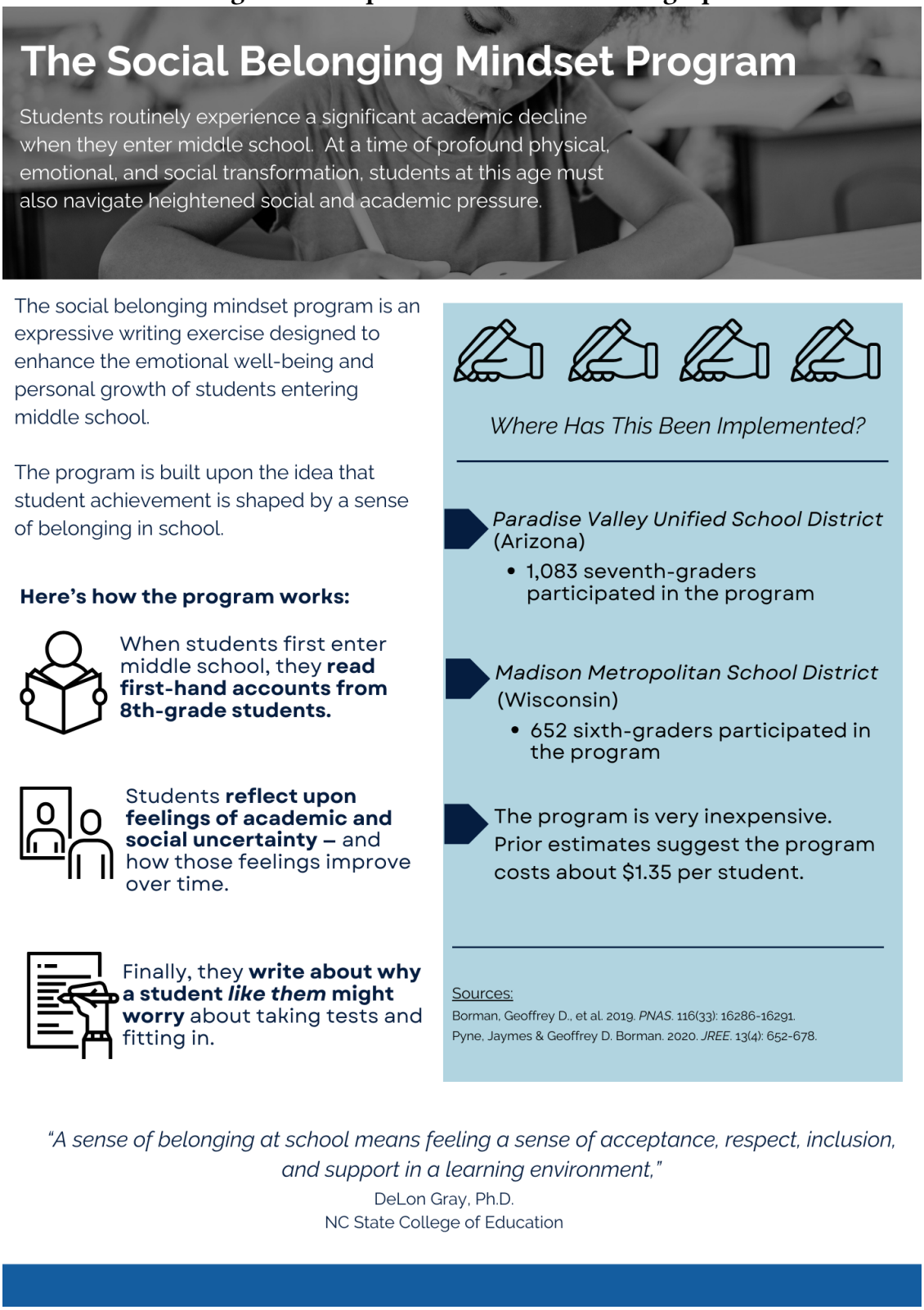
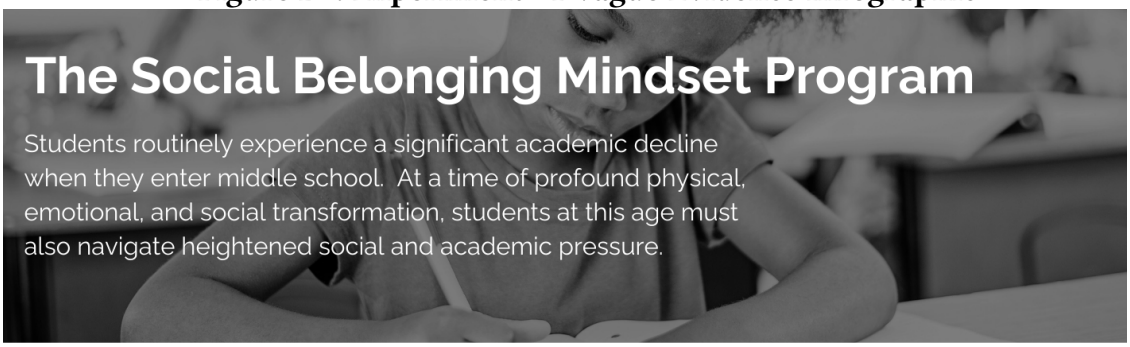


Figure B2. Experiment #1 Vague Evidence Infographic



The social belonging mindset program is an expressive writing exercise designed to enhance the emotional well-being and personal growth of students entering middle school.

The program is built upon the idea that student achievement is shaped by a sense of belonging in school.

Here's how the program works:

When students first enter middle school, they **read first-hand accounts from 8th-grade students.**

Students **reflect upon feelings of academic and social uncertainty –** and how those feelings improve over time.

Finally, they **write about why a student *like them* might worry** about taking tests and fitting in.

Where Has This Been Tested?

Paradise Valley Unified School District (Arizona)

- 1,083 seventh-graders participated in the program
- students in the program **significantly increased their GPAs**

Madison Metropolitan School District (Wisconsin)

- 652 sixth-graders participated in the program
- students in the program **significantly increased their GPAs**

The program is very inexpensive. Prior estimates suggest the program costs about $1.35 per student.

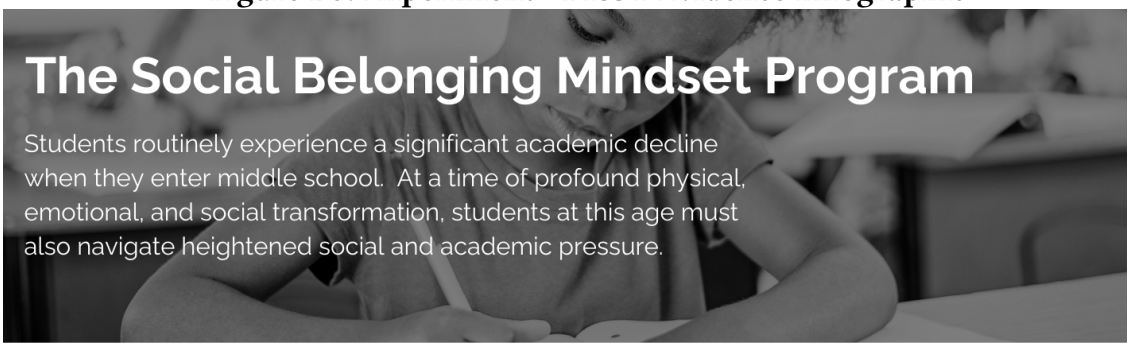
Sources:

Borman, Geoffrey D., et al. 2019. *PNAS*. 116(33): 16286-16291.
Pyne, Jaymes & Geoffrey D. Borman. 2020. *JREE*. 13(4): 652-678.

"A sense of belonging at school means feeling a sense of acceptance, respect, inclusion, and support in a learning environment,"

DeLon Gray, Ph.D.
NC State College of Education

Figure B3. Experiment #1 RCT Evidence Infographic



The social belonging mindset program is an expressive writing exercise designed to enhance the emotional well-being and personal growth of students entering middle school.

The program is built upon the idea that student achievement is shaped by a sense of belonging in school.

Here's how the program works:

When students first enter middle school, they **read first-hand accounts from 8th-grade students.**

Students **reflect upon feelings of academic and social uncertainty –** and how those feelings improve over time.

Finally, they **write about why a student *like them* might worry** about taking tests and fitting in.

"A sense of belonging at school means feeling a sense of acceptance, respect, inclusion, and support in a learning environment,"

DeLon Gray, Ph.D., NC State College of Education

Where Has This Been Tested?

Randomized Controlled Trials (RCTs)

Paradise Valley Unified School District (Arizona)

- 1,083 seventh-graders were **randomly assigned** into the program
- students in the program **significantly increased their GPAs**

Madison Metropolitan School District (Wisconsin)

- 652 sixth-graders were **randomly assigned** into the program
- students in the program **significantly increased their GPAs**

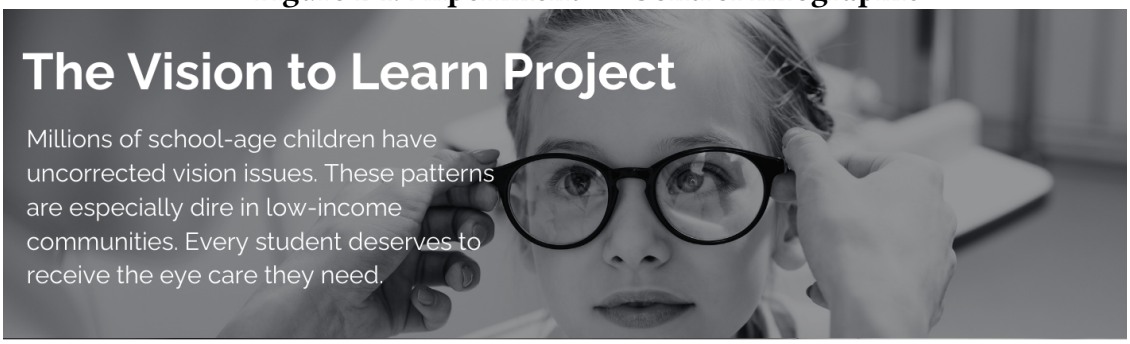
The program is very inexpensive. Prior estimates suggest the program costs about $1.35 per student.

RCTs are considered the gold standard for proving a program's effectiveness.

Random assignment ensures fair comparison groups, so researchers can be confident their results aren't due to chance or other factors.

Sources:
Borman, Geoffrey D., et al. 2019. *PNAS*. 116(33): 16286-16291.
Pyne, Jaymes & Geoffrey D. Borman. 2020. *JREE*. 13(4): 652-678.

Figure B4. Experiment #2 Control Infographic



The Vision to Learn Project tackles this issue by delivering free vision services within schools.

The program helps children overcome their vision challenges with the hope of improving each student's academic performance.

Here's how the program works:

A mobile eye clinic visits schools and conducts **free vision screenings and eye examinations.**

If a student needs glasses, they get to **choose frames they feel comfortable and confident in.**

Students are given **free glasses** — including **replacements** — for one year.

Where Has This Been Implemented?

Baltimore City Public Schools (Maryland)

- 2,304 students participated in the Vision to Learn Project.

The program helped many students receive vision care for the first time in their lives.

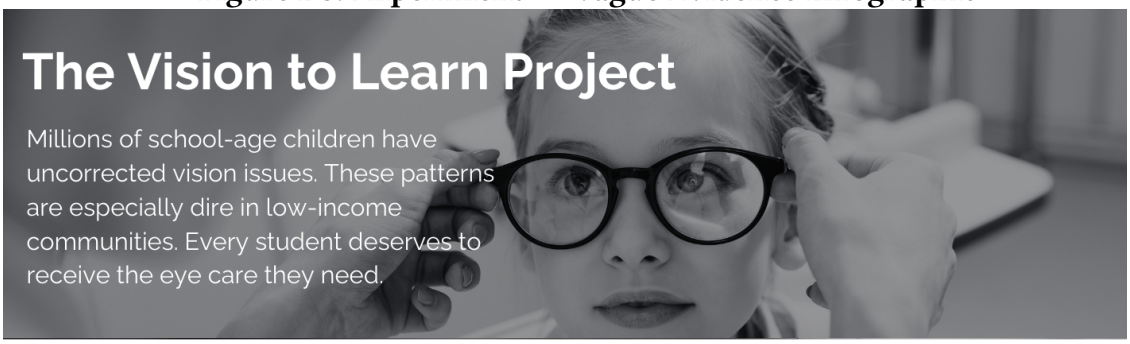
Source:

Neitzel, Amanda J., et al. 2021. *JAMA Ophthalmology* 139(10): 1104-1114.

"We know, based on common sense, that giving glasses for kids is important for education [and] health,"

Dr. Leana Wen,
Baltimore City Health Commissioner

Figure B5. Experiment #2 Vague Evidence Infographic



The Vision to Learn Project tackles this issue by delivering free vision services within schools.

The program helps children overcome their vision challenges with the hope of improving each student's academic performance.

Here's how the program works:

A mobile eye clinic visits schools and conducts **free vision screenings and eye examinations.**

If a student needs glasses, they get to **choose frames they feel comfortable and confident in.**

Students are given **free glasses** — including **replacements** — for one year.

Where Has This Been Tested?

Baltimore City Public Schools (Maryland)

- 2,304 students participated in the Vision to Learn Project.

After two years, researchers:

- **did not find statistically significant effects** on state-mandated math & reading scores.
- **lacked conclusive evidence** that the program had an enduring impact on student learning.

Nevertheless, the program helped many students receive vision care for the first time in their lives.

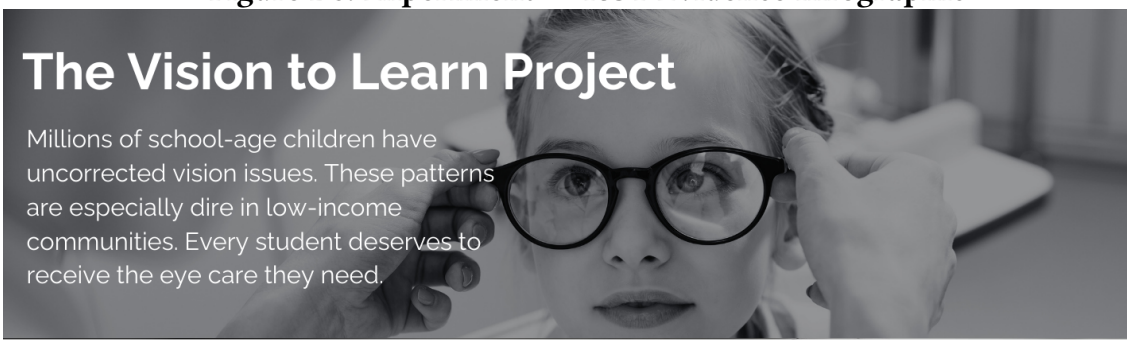
Source:

Neitzel, Amanda J., et al. 2021. *JAMA Ophthalmology* 139(10): 1104-1114.

"We know, based on common sense, that giving glasses for kids is important for education [and] health,"

Dr. Leana Wen,
Baltimore City Health Commissioner

Figure B6. Experiment #2 RCT Evidence Infographic



The Vision to Learn Project tackles this issue by delivering free vision services within schools.

The program helps children overcome their vision challenges with the hope of improving each student's academic performance.

Here's how the program works:

A mobile eye clinic visits schools and conducts **free vision screenings and eye examinations.**

If a student needs glasses, they get to **choose frames they feel comfortable and confident in**.

Students are given **free glasses** — including **replacements** — for one year.

"We know, based on common sense, that giving glasses for kids is important for education [and] health,"

Dr. Leana Wen,
Baltimore City Health Commissioner

Randomized Controlled Trials (RCTs)

Where Has This Been Tested?

Baltimore City Public Schools (Maryland)

- 2,304 students participated in a **randomized controlled trial to evaluate the impact** of the Vision to Learn Project.

After two years, researchers:

- **did not find statistically significant effects** on state-mandated math & reading scores.
- **lacked conclusive evidence** that the program had an enduring impact on student learning.

Nevertheless, the program helped many students receive vision care for the first time in their lives.

RCTs are considered the gold standard for proving a program's effectiveness.

Random assignment ensures fair comparison groups, so researchers can be confident their results aren't due to chance or other factors.

Source:

Neitzel, Amanda J., et al. 2021. *JAMA Ophthalmology* 139(10): 1104-1114.

Figure B7. Experiment #3 Control Proposal

From: The Lafayette College Gov Lab
Date: October 1, 2023
Subject: Exploring New Generative AI Tutoring Services
To: [Recipient's Name]

Executive Summary
This proposal aims to explore **generative AI tutoring solutions** in American classrooms.

Background
Recent advancements in AI have ushered in generative AI tutoring systems that offer personalized and scalable educational solutions. These systems adapt to the unique needs of individual learners, providing tailored guidance across various subjects. For example, Quizlet Plus allows students to ask questions to an AI tutor (Q-Chat), generate subject-specific flashcards, and work through customized practice quizzes.

Benefits and Challenges Addressed
- Personalization: AI-based tutoring customizes instruction to meet diverse learning needs, enhancing the overall student experience.
- Scalability: Generative AI tutoring may be as inexpensive as an annual fee of $36 per student, providing a cost-effective means of reaching a large student population in need of tutoring services.
- Data-Driven Decisions: Real-time data on student progress enables transparent and informed decision-making for educators.
- Consistency: Standardized AI-driven instruction ensures uniform quality across educational modules.

Thank you for considering our **proposal to explore new, generative AI tutoring services**. We believe that this emerging technology presents an innovative opportunity to meet a critically important need for scalable, personalized, high-quality tutoring services. Should you have any questions or require additional information, please don't hesitate to reach out (govlab@lafayette.edu).

Students of the Lafayette College Gov Lab

Note: We redacted any potentially author-identifying information.

Figure B8. Experiment #3 RCT Pilot Proposal

From: The Lafayette College Gov Lab
Date: October 1, 2023
Subject: Proposing an RCT Pilot to Explore New Generative AI Tutoring Services
To: [Recipient's Name]

Executive Summary
This proposal aims to explore **generative AI tutoring solutions** in American classrooms. By implementing **a Randomized Controlled Trial (RCT) pilot program**, schools can rigorously evaluate the costs and benefits of this emerging technology.

Background
Recent advancements in AI have ushered in generative AI tutoring systems that offer personalized and scalable educational solutions. These systems adapt to the unique needs of individual learners, providing tailored guidance across various subjects. For example, Quizlet Plus allows students to ask questions to an AI tutor (Q-Chat), generate subject-specific flashcards, and work through customized practice quizzes.

Benefits and Challenges Addressed
- Personalization: AI-based tutoring customizes instruction to meet diverse learning needs, enhancing the overall student experience.
- Scalability: Generative AI tutoring may be as inexpensive as an annual fee of $36 per student, providing a cost-effective means of reaching a large student population in need of tutoring services.
- Data-Driven Decisions: Real-time data on student progress enables transparent and informed decision-making for educators.
- Consistency: Standardized AI-driven instruction ensures uniform quality across educational modules.

The Need for a Randomized Controlled Trial (RCT) Pilot Program
- ***RCTs are considered the gold standard for proving a program's effectiveness*** because they create fair and unbiased comparison groups.
- ***An RCT pilot program would allow schools to rigorously evaluate if generative AI tutoring works*** *before* committing the financial resources necessary to offer these services to all students.

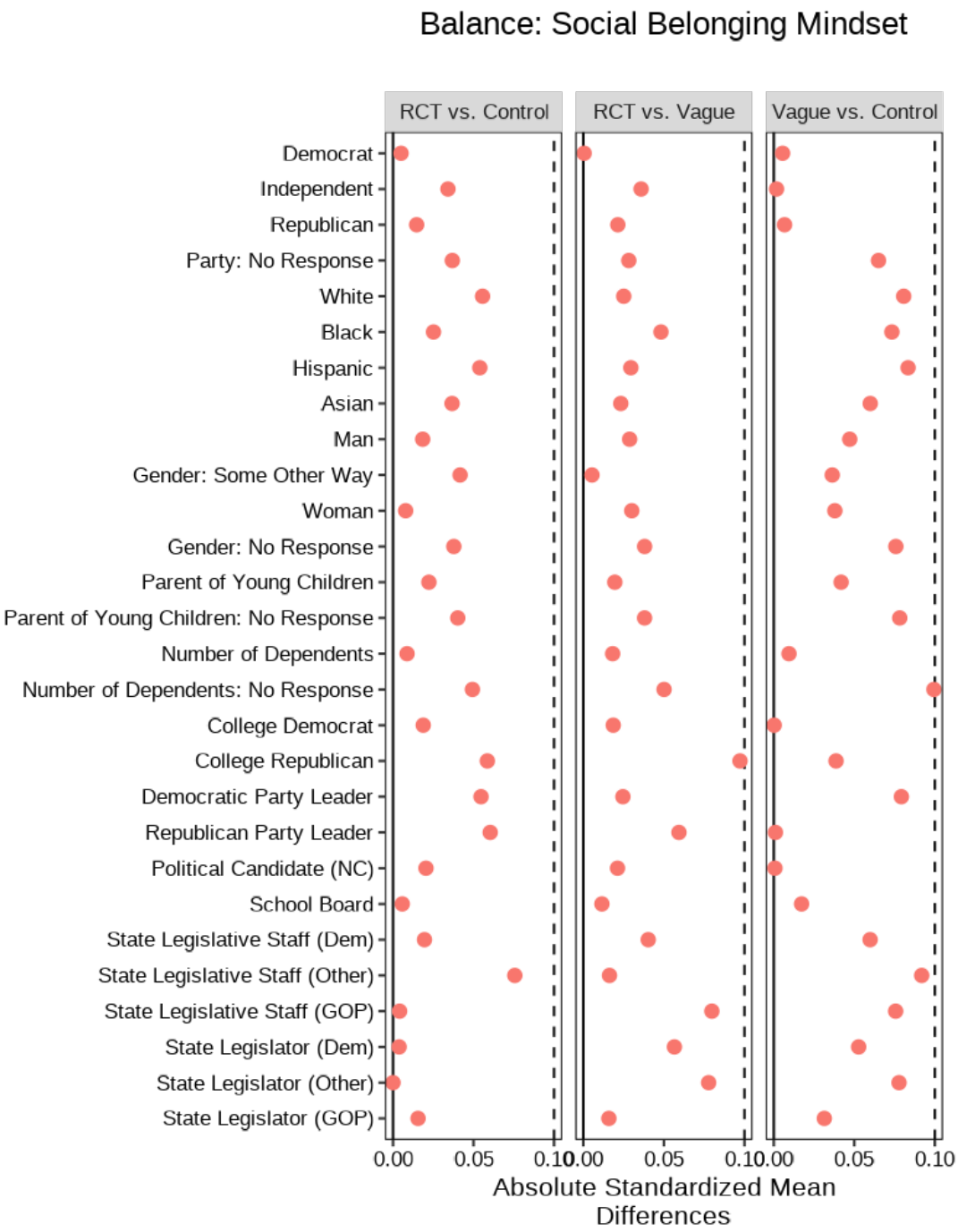
Thank you for considering our **proposed RCT pilot program to evaluate new, generative AI tutoring services**. We believe that this emerging technology presents an innovative opportunity to meet a critically important need for scalable, personalized, high-quality tutoring services. Should you have any questions or require additional information, please don't hesitate to reach out (govlab@lafayette.edu).

Students of the Lafayette College Gov Lab

Note: We redacted any potentially author-identifying information.

C Additional Analyses

Figure C1. Covariate Balance, Social Belonging Mindset Experiment



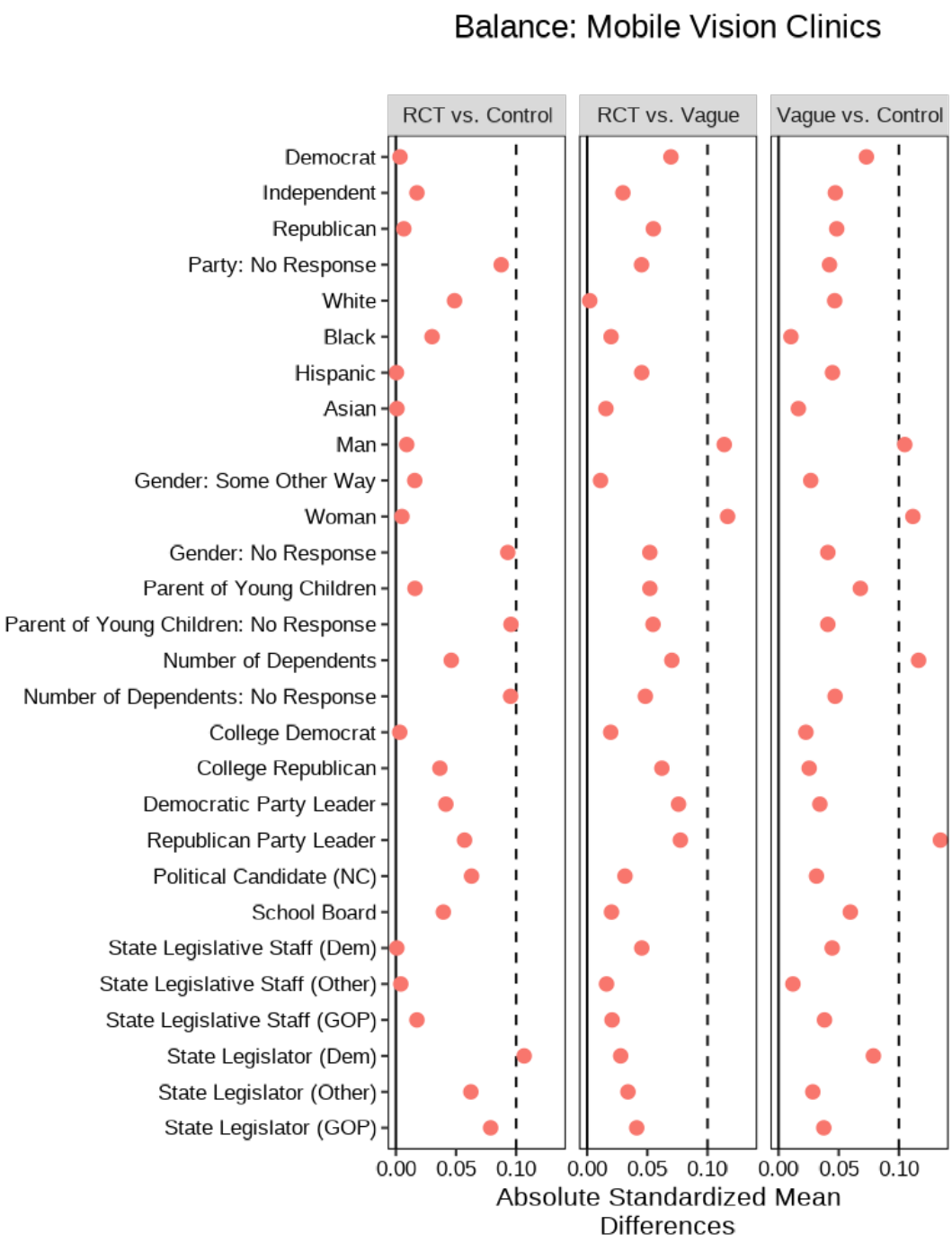
Note: Points are absolute standardized mean differences (pooled standard deviation) in respondent characteristics between treatment arms for the social belonging mindset experiment. The dashed line marks 0.10, a standard benchmark. Points near or beyond the line are not by themselves evidence of imbalance. Overall the arms are well balanced: no point exceeds 0.10 (the largest is 0.099).

Table C1. Covariate Balance, Social Belonging Mindset Experiment

	Standardized Mean Difference		
	Vague-Control	RCT-Control	RCT-Vague
Democrat	0.005	0.005	0.000
Independent	0.002	−0.034	−0.036
Republican	−0.007	0.015	0.021
Party: No Response	−0.065	−0.037	0.028
White	0.081	0.056	−0.025
Black	−0.073	−0.025	0.048
Hispanic	0.083	0.054	−0.029
Asian	0.060	0.037	−0.023
Man	0.047	0.018	−0.029
Gender: Some Other Way	−0.036	−0.042	−0.005
Woman	−0.038	−0.008	0.030
Gender: No Response	−0.076	−0.038	0.038
Parent of Young Children	0.042	0.022	−0.020
Parent of Young Children: No Response	−0.078	−0.040	0.038
Number of Dependents	−0.009	0.009	0.018
Number of Dependents: No Response	−0.099	−0.049	0.050
College Democrat	0.000	−0.019	−0.019
College Republican	0.039	−0.059	−0.097
Democratic Party Leader	0.079	0.055	−0.025
Republican Party Leader	−0.001	−0.060	−0.059
Political Candidate (NC)	−0.001	0.020	0.021
School Board	−0.017	−0.006	0.012
State Legislative Staff (Dem)	0.060	0.020	−0.040
State Legislative Staff (Other)	−0.092	−0.076	0.016
State Legislative Staff (GOP)	−0.076	0.004	0.080
State Legislator (Dem)	−0.053	0.004	0.057
State Legislator (Other)	0.078	0.000	−0.078
State Legislator (GOP)	−0.031	−0.016	0.016

Note: Cell entries are standardized mean differences in respondent characteristics between treatment arms (pooled standard deviation). Shading marks every cell in a row that has at least one comparison above |0.10|. We believe the arms are well balanced on observables, as no comparison exceeds |0.10| (the largest is 0.099), so no row is shaded. That said, the 0.10 benchmark is a conservative benchmark, not a statistical test.

Figure C2. Covariate Balance, Mobile Vision Clinic Experiment



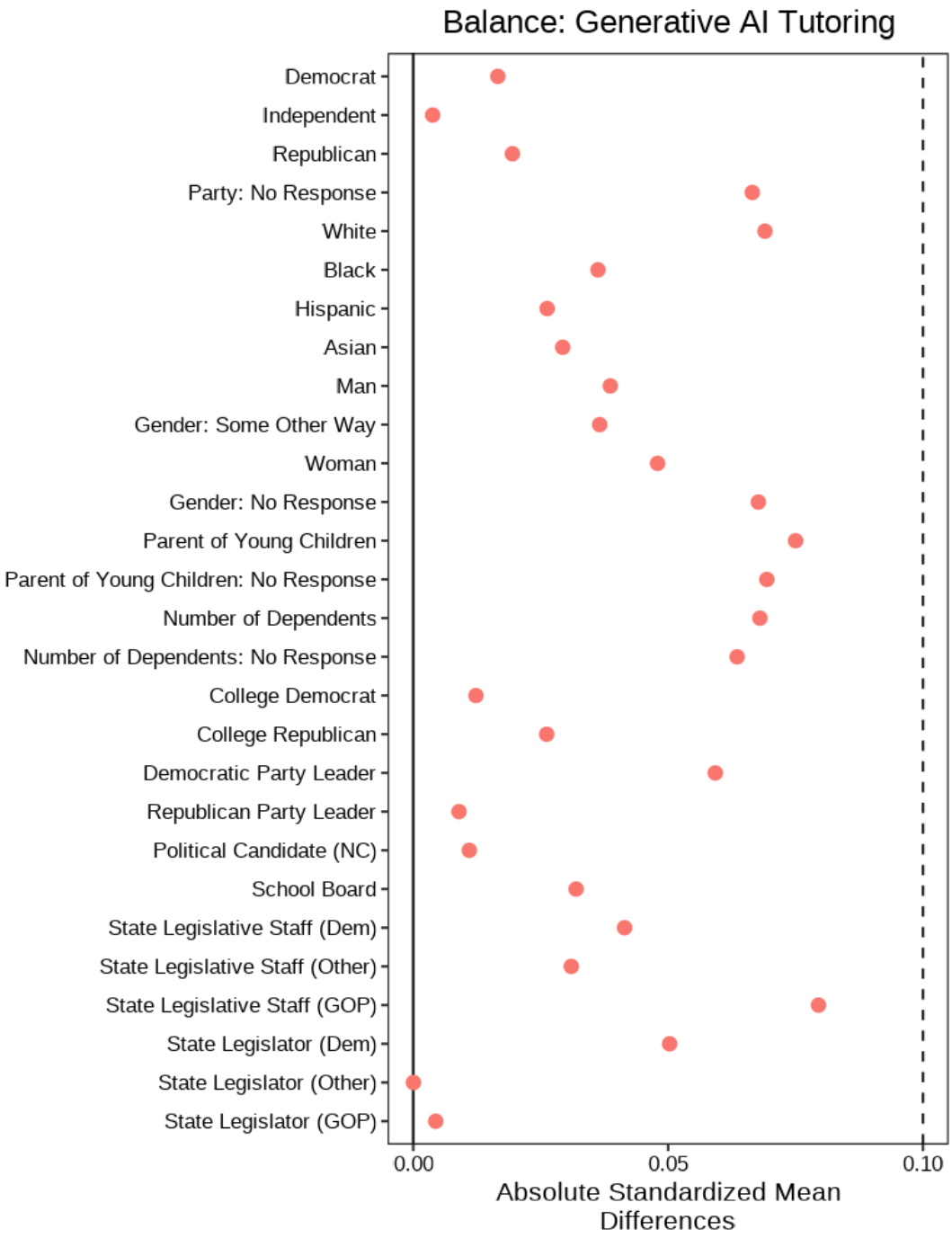
Note: Points are absolute standardized mean differences (pooled standard deviation) in respondent characteristics between treatment arms for the mobile vision clinic experiment. The dashed line marks 0.10, which should be interpreted as a rough benchmark, rather than a clear statistical test. Points near or beyond the line are not by themselves evidence of imbalance. Overall the arms are well balanced: only seven comparisons exceed 0.10 (with the largest only 0.135).

Table C2. Covariate Balance, Mobile Vision Clinic Experiment

	Standardized Mean Difference		
	Vague-Control	RCT-Control	RCT-Vague
Democrat	0.073	0.003	−0.070
Independent	−0.047	−0.018	0.030
Republican	−0.048	0.007	0.055
Party: No Response	0.042	0.088	0.045
White	−0.047	−0.049	−0.002
Black	−0.010	−0.030	−0.020
Hispanic	0.045	0.000	−0.045
Asian	−0.016	−0.001	0.016
Man	−0.105	0.009	0.114
Gender: Some Other Way	−0.027	−0.016	0.011
Woman	0.112	−0.005	−0.117
Gender: No Response	0.041	0.093	0.052
Parent of Young Children	−0.068	−0.016	0.052
Parent of Young Children: No Response	0.041	0.096	0.055
Number of Dependents	−0.116	−0.046	0.070
Number of Dependents: No Response	0.047	0.095	0.048
College Democrat	0.023	0.003	−0.020
College Republican	0.025	−0.037	−0.062
Democratic Party Leader	0.034	−0.042	−0.076
Republican Party Leader	−0.135	−0.057	0.077
Political Candidate (NC)	0.032	0.063	0.031
School Board	−0.060	−0.039	0.020
State Legislative Staff (Dem)	0.045	−0.001	−0.045
State Legislative Staff (Other)	−0.012	0.004	0.016
State Legislative Staff (GOP)	0.038	0.018	−0.021
State Legislator (Dem)	0.079	0.107	0.028
State Legislator (Other)	−0.028	−0.062	−0.034
State Legislator (GOP)	0.038	0.079	0.041

Note: Cell entries are standardized mean differences in respondent characteristics between treatment arms (pooled standard deviation). Shading marks every cell in a row that has at least one comparison above |0.10|. We believe the arms are well balanced on observables, as only seven of the many pairwise comparisons across the three columns exceed |0.10| – and the largest of these is |0.135|.

Figure C3. Covariate Balance, Generative AI Tutoring Experiment



Note: Points are absolute standardized mean differences (pooled standard deviation) in respondent characteristics between treatment arms for the generative AI tutoring experiment. The dashed line marks 0.10, one rough benchmark (not a firm statistical test). Points near or beyond the line are not by themselves evidence of imbalance. Overall the arms are well balanced: no point exceeds 0.10 (and the largest is 0.080).

Table C3. Covariate Balance, Generative AI Tutoring Experiment

	Standardized Mean Difference
	Treatment-Control
Democrat	0.017
Independent	0.004
Republican	−0.019
Party: No Response	0.067
White	−0.069
Black	−0.036
Hispanic	0.026
Asian	−0.029
Man	−0.039
Gender: Some Other Way	−0.037
Woman	0.048
Gender: No Response	0.068
Parent of Young Children	0.075
Parent of Young Children: No Response	0.069
Number of Dependents	0.068
Number of Dependents: No Response	0.064
College Democrat	0.012
College Republican	0.026
Democratic Party Leader	0.059
Republican Party Leader	0.009
Political Candidate (NC)	0.011
School Board	0.032
State Legislative Staff (Dem)	−0.041
State Legislative Staff (Other)	−0.031
State Legislative Staff (GOP)	−0.080
State Legislator (Dem)	−0.050
State Legislator (Other)	0.000
State Legislator (GOP)	−0.004

Note: Cell entries are standardized mean differences in respondent characteristics between treatment arms (pooled standard deviation). Shading marks every cell in a row that has at least one comparison above |0.10| and is a visual aid only. We argue that the arms are well balanced on observables, as no comparison exceeds |0.10| (the largest is 0.080). It is worth noting that the 0.10 benchmark is a rough benchmark, not a statistical test.

Table C4. Strata-FE Regression Results for Experiment #1, *Excluding* College Partisans

	Social Belonging Mindset Program (Positive Results)		
	Program Effectiveness (0–100)	Program Support (0–100)	Program Support (0,1)
Experimental Evidence Treatment	9.407***	5.507***	0.074***
	(1.217)	(1.179)	(0.020)
Vague Evidence Treatment	8.627***	5.416***	0.079***
	(1.203)	(1.145)	(0.020)
Strata Fixed Effects	✓	✓	✓
N	2,566	2,711	2,711

Note: Cell entries are OLS coefficients with robust (HC2) standard errors in parentheses. Models include strata fixed effects. College Democrats and College Republicans are excluded. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Table C5. Strata-FE Regression Results for Experiment #2, *Excluding* College Partisans

	Mobile Vision Clinic (Null Results)		
	Program Effectiveness (0–100)	Program Support (0–100)	Program Support (0,1)
Experimental Evidence Treatment	−17.136***	−4.531***	−0.043**
	(1.244)	(1.059)	(0.016)
Vague Evidence Treatment	−18.333***	−4.023***	−0.037*
	(1.242)	(1.016)	(0.016)
Strata Fixed Effects	✓	✓	✓
N	2,607	2,697	2,697

Note: Cell entries are OLS coefficients with robust (HC2) standard errors in parentheses. Models include strata fixed effects. College Democrats and College Republicans are excluded. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Table C6. Strata-FE Regression Results for Experiment #3, *Excluding* College Partisans

	Generative AI Tutoring (RCT Pilot Results)	
	Program Support (0–100)	Program Support (0,1)
RCT Pilot Treatment	−0.673	−0.011
	(1.002)	(0.019)
Strata Fixed Effects	✓	✓
N	2,646	2,646

Note: Cell entries are OLS coefficients with robust (HC2) standard errors in parentheses. Models include strata fixed effects. College Democrats and College Republicans are excluded. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Table C7. Strata-FE Regression Results for Experiment #1, Elected Board of Education Members, State Lawmakers, and Legislative Staffers *Only*

	Social Belonging Mindset Program (Positive Results)		
	Program Effectiveness (0–100)	Program Support (0–100)	Program Support (0,1)
Experimental Evidence Treatment	9.199***	5.895***	0.079***
	(1.363)	(1.324)	(0.022)
Vague Evidence Treatment	7.939***	5.239***	0.073**
	(1.346)	(1.295)	(0.022)
Strata Fixed Effects	✓	✓	✓
N	2,055	2,172	2,172

Note: Cell entries are OLS coefficients with robust (HC2) standard errors in parentheses. Models include strata fixed effects. Sample restricted to elected officials and legislative staffers. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Table C8. Strata-FE Regression Results for Experiment #2, Elected Board of Education Members, State Lawmakers, and Legislative Staffers *Only*

	Mobile Vision Clinic (Null Results)		
	Program Effectiveness (0–100)	Program Support (0–100)	Program Support (0,1)
Experimental Evidence Treatment	−17.406***	−4.671***	−0.044*
	(1.396)	(1.186)	(0.018)
Vague Evidence Treatment	−17.506***	−3.366**	−0.021
	(1.399)	(1.144)	(0.017)
Strata Fixed Effects	✓	✓	✓
N	2,082	2,155	2,155

Note: Cell entries are OLS coefficients with robust (HC2) standard errors in parentheses. Models include strata fixed effects. Sample restricted to elected officials and legislative staffers. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Table C9. Strata-FE Regression Results for Experiment #3, Elected Board of Education Members, State Lawmakers, and Legislative Staffers *Only*

	Generative AI Tutoring (RCT Pilot Results)	
	Program Support (0–100)	Program Support (0,1)
RCT Pilot Treatment	−1.049	−0.024
	(1.111)	(0.022)
Strata Fixed Effects	✓	✓
N	2,119	2,119

Note: Cell entries are OLS coefficients with robust (HC2) standard errors in parentheses. Models include strata fixed effects. Sample restricted to elected officials and legislative staffers. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Table C10. Regression Results for Experiment #1 *Without* Strata Fixed Effects

	Social Belonging Mindset Program (Positive Results)		
	Program Effectiveness (0–100)	Program Support (0–100)	Program Support (0,1)
Experimental Evidence Treatment	9.913***	6.116***	0.084***
	(1.229)	(1.209)	(0.020)
Vague Evidence Treatment	9.055***	5.929***	0.084***
	(1.219)	(1.198)	(0.020)
N	2,582	2,727	2,727

Note: Cell entries are OLS coefficients with conventional (non-robust) standard errors in parentheses. Models do not include strata fixed effects. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Table C11. Regression Results for Experiment #2 *Without* Strata Fixed Effects

	Mobile Vision Clinic (Null Results)		
	Program Effectiveness (0–100)	Program Support (0–100)	Program Support (0,1)
Experimental Evidence Treatment	−17.093***	−4.504***	−0.044**
	(1.306)	(1.080)	(0.016)
Vague Evidence Treatment	−17.787***	−3.374**	−0.028
	(1.309)	(1.081)	(0.016)
N	2,622	2,715	2,715

Note: Cell entries are OLS coefficients with conventional (non-robust) standard errors in parentheses. Models do not include strata fixed effects. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Table C12. Regression Results for Experiment #3 *Without* Strata Fixed Effects

	Generative AI Tutoring (RCT Pilot Results)	
	Program Support (0–100)	Program Support (0,1)
RCT Pilot Treatment	−0.535	−0.008
	(1.003)	(0.019)
N	2,662	2,662

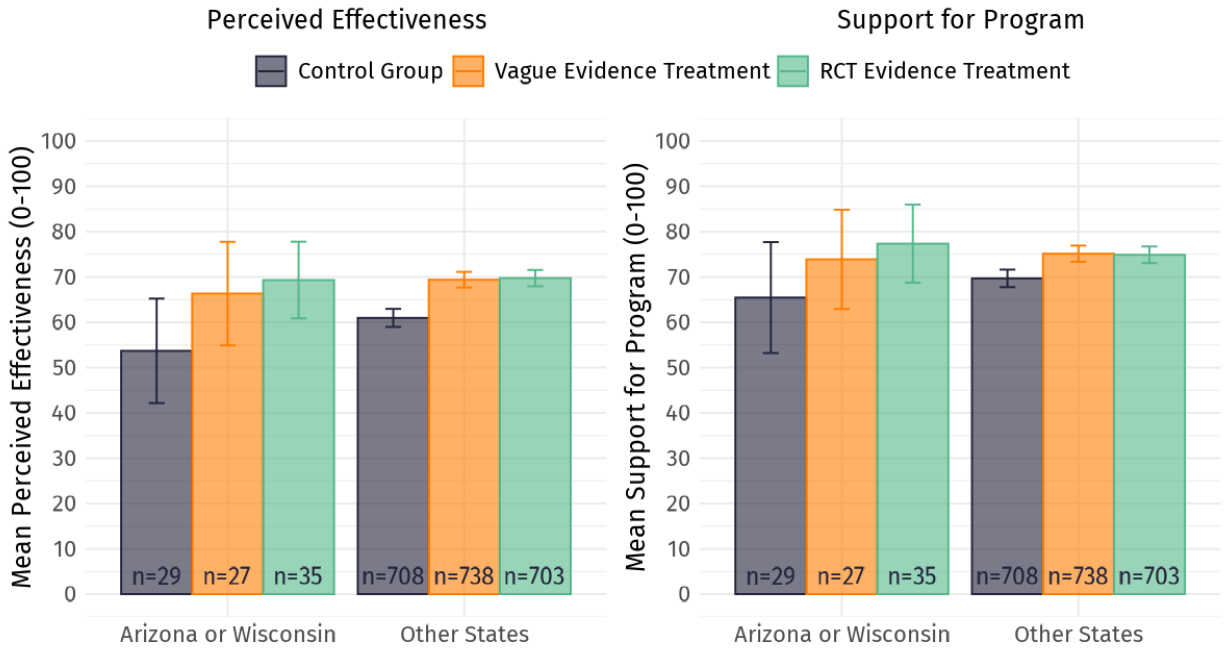
Note: Cell entries are OLS coefficients with conventional (non-robust) standard errors in parentheses. Models do not include strata fixed effects. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Table C13. Randomization Inference Results for All Experiments

Contrast	Coefficient	RI p-value	N	N Dropped	Strata
Social Belonging: Effectiveness (0–100)					
RCT vs. Control	9.292	<0.001	1,702	3	10
Vague vs. Control	8.675	<0.001	1,733	1	11
RCT vs. Vague	0.794	0.487	1,719	6	10
Social Belonging: Support (0–100)					
RCT vs. Control	5.368	<0.001	1,800	3	10
Vague vs. Control	5.397	<0.001	1,835	1	11
RCT vs. Vague	0.086	0.938	1,809	6	10
Social Belonging: Support (0,1)					
RCT vs. Control	0.072	<0.001	1,800	3	10
Vague vs. Control	0.078	<0.001	1,835	1	11
RCT vs. Vague	−0.005	0.788	1,809	6	10
Mobile Vision Clinics: Effectiveness (0–100)					
RCT vs. Control	−17.122	<0.001	1,777	1	11
Vague vs. Control	−18.266	<0.001	1,769	1	11
RCT vs. Vague	1.061	0.470	1,696	0	11
Mobile Vision Clinics: Support (0–100)					
RCT vs. Control	−4.498	<0.001	1,836	1	11
Vague vs. Control	−4.023	<0.001	1,834	0	12
RCT vs. Vague	−0.480	0.674	1,758	1	11
Mobile Vision Clinics: Support (0,1)					
RCT vs. Control	−0.044	0.007	1,836	1	11
Vague vs. Control	−0.037	0.021	1,834	0	12
RCT vs. Vague	−0.008	0.646	1,758	1	11
Generative AI Tutoring: Support (0–100)					
RCT Pilot vs. No Pilot	−0.556	0.604	2,662	0	12
Generative AI Tutoring: Support (0,1)					
RCT Pilot vs. No Pilot	−0.010	0.633	2,662	0	12

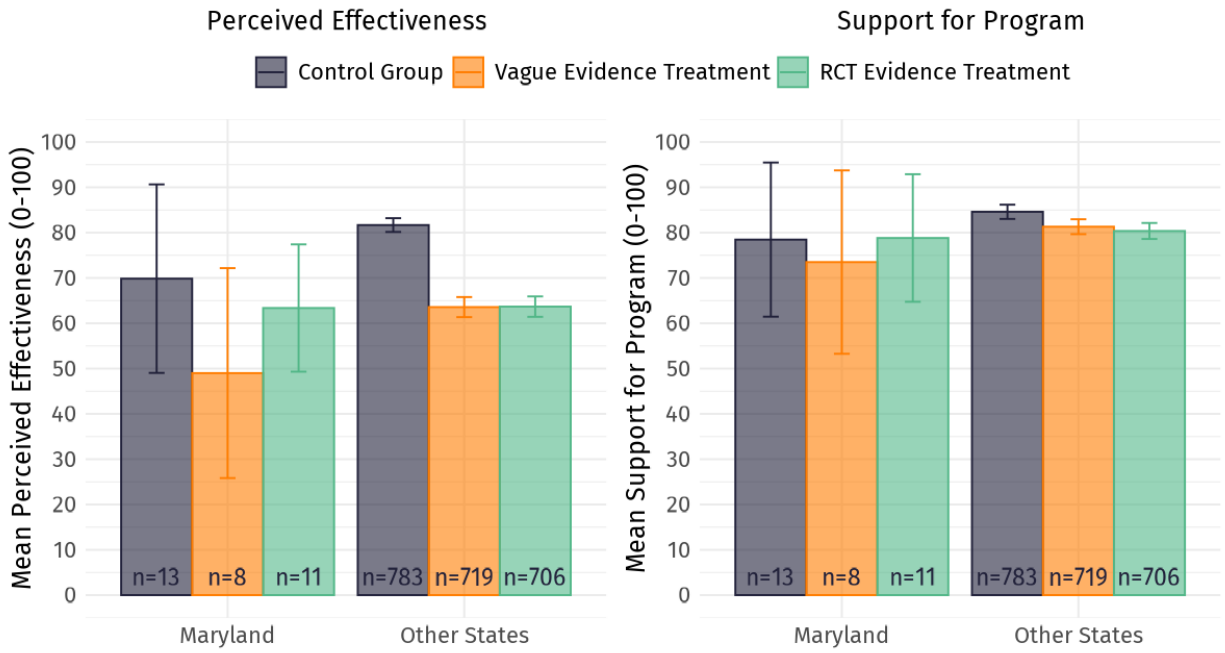
Note: Coefficients are stratified differences in means (each stratum weighted by its number of respondents), estimated separately for each pairwise contrast; they can therefore differ slightly from the OLS strata fixed-effects coefficients reported in the main text. RI p-values are two-sided, from 1,000 re-randomizations under the sharp null of no effect for any respondent, holding the observed number of treated respondents in each stratum fixed. Strata are the 12 sub-samples. A stratum is dropped from an analysis when all of its respondents (or its only respondent) fall in one group, because it contains no treated-versus-comparison variation and cannot contribute to the estimate; the number of strata therefore differs across contrasts and outcomes, and N Dropped is the number of respondents in dropped strata (either non-partisan state legislators, College Republicans, or both). Randomization inference was not pre-registered.

Figure C4. Support and Effectiveness by In-State Status, Social Belonging Mindset Program



Note: Bars are mean perceived effectiveness (left) and mean support for the program (right), 0–100, with 95% confidence intervals; cell sizes are printed at the base of each bar. Respondents from Arizona or Wisconsin, the states named in the social belonging infographics, are compared with respondents from all other states. Descriptive only; no tests.

Figure C5. Support and Effectiveness by In-State Status, Mobile Vision Clinic



Note: Bars are mean perceived effectiveness (left) and mean support for the program (right), 0–100, with 95% confidence intervals; cell sizes are printed at the base of each bar. Respondents from Maryland, the state named in the mobile vision clinic infographics, are compared with respondents from all other states. Descriptive only; no tests.