Randomized Policy Experiments Increase Legitimacy

Aaron Martin[†]**, Kyle Peyton**[‡]

Abstract. Randomized experiments are widely promoted in science and industry for their ability to provide credible evidence about the effects of costly interventions prior to large-scale implementations. In business and medicine, for example, randomized experiments consistently inform high-stakes decisions that promote economic growth and human welfare. Despite these potential benefits, governments regularly implement new policies on large populations without credible evidence about their potential benefits or harms. Moreover, there is rarely a mechanism in place that allows for a credible evaluation of the effects of untested policies after they have been implemented. This practice has profound implications for social welfare and raises normative concerns about the legitimacy of government decision-making. Here, we investigate whether the scarcity of randomized policy experiments might partly be explained by public opposition to experimentation. Across two studies sampling the Australian adult population, we find that the vast majority of individuals prefer randomized policy experiments prior to the implementation of campaign promises. Conducting these experiments significantly increases perceptions of government legitimacy across multiple social policy domains, and governments are not penalized when an experiment shows that a policy has a negative impact. These findings demonstrate that voters are not averse to experimentation and suggest that governments would benefit from an increased focus on randomized policy experiments.

1 Overview

Randomized experiments are widely promoted in science and industry for their ability to provide credible evidence about the effects of costly interventions before large-scale implementations. In fields like business and medicine, randomized experiments consistently inform high-stakes decisions that promote economic growth and human welfare. However, despite these potential benefits, governments frequently implement new policies on large populations without credible evidence of their potential impacts. This practice undermines social welfare and raises significant normative concerns about the legitimacy of government decision-making.

Implementing policies without randomized trials leads to significant opportunity costs. Every non-randomized policy implementation represents a missed opportunity to gather valuable insights into what works and what doesn't. Subsequent claims about the effects of non-randomized policy experiments are inherently questionable and potentially misleading (Gerber, Green and Kaplan, 2014). The inability to learn from non-randomized implementations negatively affects social welfare by preventing the accumulation of reliable evidence needed to inform future policy decisions.

The status quo of government policy implementation suffers from normative issues such as arbitrariness, irrevisability, and lack of public justification (Tanasoca and Leigh, 2024). Randomized trials, by contrast, inherently possess virtues that address these issues, making them normatively superior. For

[†] School of Social and Political Sciences, University of Melbourne. aaron.martin@unimelb.edu.au..

[‡] Faculty of Business and Economics, University of Melbourne. kyle.peyton@unimelb.edu.au..

instance, randomized trials ensure non-arbitrariness through random assignment, provide a mechanism for revisability by systematically collecting evidence to inform policy adjustments, and enhance public justification by offering transparent and credible evidence of policy impacts. This raises the possibility that randomized policy experiments may not only address these normative concerns but also increase public trust and the perceived legitimacy of government decision-making.

To date, little is known about public attitudes toward policy experimentation or whether government decision-making informed by randomized experiments is viewed as legitimate. Recent empirical work suggests that people may prefer the universal implementation of untested policies over randomized trials, potentially posing a barrier to evidence-based policymaking (Meyer et al., 2019; Heck et al., 2020). However, this work, which draws primarily on surveys about experimentation in the medical and corporate sectors, has reached mixed conclusions (Mislavsky, Dietvorst and Simonsohn, 2020; Mazar, Elbaek and Mitkidis, 2023).

There are several reasons why people might prefer the immediate implementation of untested policies over randomized experiments. First, risk aversion can lead individuals to favor the status quo or no action when policy effects are uncertain, perceiving experimentation as a greater risk. Second, the fear of negative outcomes from trials could result in the abandonment of favored policies, discouraging support for experimentation. Finally, people may be biased towards implementing policies they like, prioritizing personal or ideological support over empirical evidence of efficacy.

Here, we investigate whether the scarcity of randomized policy experiments might partly be explained by public opposition to experimentation. Across two studies sampling the Australian adult population, we find that the vast majority of individuals prefer randomized policy experiments before the implementation of campaign promises. Conducting these experiments significantly increases perceptions of government legitimacy across multiple policy domains, and governments are not penalized when an experiment shows that a policy has a negative impact.

Our findings demonstrate that voters are not averse to experimentation and suggest that governments would benefit from an increased focus on randomized policy experiments. By embracing a scientific approach to policy implementation, governments can justify their actions with credible evidence and mitigate normative concerns raised by the status quo practice of implementing untested policies on large populations without credible evidence of their potential benefits or harms.

2 Experimental Design

To examine public preferences for randomized policy experiments, and whether governments that conduct them are viewed as more legitimate, we conducted two randomized survey experiments.

In Experiment 1, we randomly assigned respondents to scenarios where a recently elected government decided to either implement a policy from a campaign promise, conduct a randomized trial before implementation, or abandon the policy. This design allows us to examine the effects of a government decision to conduct a randomized policy experiment or abandon a campaign promise, relative to the status quo of implementing the policy untested.

In Experiment 2, respondents were informed that the government had decided to conduct a randomized trial to evaluate their policy prior to a nationwide implementation. We then randomly assigned them to scenarios where the experiment revealed that the policy had a positive impact, negative impact, or no impact. This design enables us to examine whether, given a government decision to conduct an experiment, perceived legitimacy is affected by whether the results demonstrate a positive or negative impact of the policy.

Section 2.1 provides an overview of the sample, and Section 2.2 describes the experimental designs in more detail. Additional details are provided in the Supporting Information (SI) Appendix. The pre-registration is available at aspredicted.org/FD9_FV7.

2.1 Sample

We contracted with Qualtrics Panels to recruit a sample of 1,600 Australian adults (aged 18+) to yield a nationally representative sample on census marginals (e.g., age, gender). Sample size was determined using simulation-based power calculations targeting a minimum detectable effect of 0.20 standard units with at least 90% power for null hypothesis tests and equivalence tests involving pairwise comparisons between treatment arms. The survey was fielded between 28 February and 21 March 2024, and resulted in a sample of 1,782 respondents that consented to take the survey and passed pre-treatment attention check questions. Figure S6 provides an overview of the survey design and randomization. Table S3 provides summary statistics for demographic characteristics, with comparisons to census marginals where available.

2.1.1 Pre-treatment trust in government

We measured pre-treatment trust in government after the demographic questions using the following four items, each with a 7-point scale from "Strongly disagree" to "Strongly agree".[1]

1. "I am comfortable allowing the government to decide how best to deal with social problems."
2. "I am confident in the government's ability to make decisions about policy implementation when there's uncertainty about policy outcomes."
3. "The government is generally truthful in its communications with the Australian public."
4. "The government typically follows through on its commitments and promises."

Responses were highly correlated ($\alpha = 0.91$), and we constructed a summary index of *Trust in*

[1] Question prompt: "Please say whether you agree or disagree with the following statements about the Australian Government. When answering the questions, please consider the Australian Government in general, rather than the current government or any specific political party."

Government using inverse covariance weighting (Anderson, 2008). The average level of *Trust in Government* in the sample was 3.75 (sd: 1.44; range: 1 to 7). Supporters of the Labor Party reported the highest levels (mean: 4.32; sd: 1.27) and those indicating they were not a supporter of any political party reported the lowest levels (mean: 2.47; sd: 1.35).

2.2 Design

After completing background survey questions, respondents read a short educational primer on randomized experiments and answered three comprehension questions (see SI for additional details). The goal of this primer was to ensure a common understanding of the logic of randomized experiments and key concepts such as "random assignment".[2] Participants were subsequently randomly assigned to one of two parallel experiments.

Experiment 1 ($n = 898$ respondents $\times 3$ vignettes = 2,694 observations): Respondents were assigned to one of three conditions with equal probability for each policy scenario: 1) Implement: proceed with the policy as planned without an experiment; 2) Experiment: conduct a randomized experiment before deciding on implementation; 3) Neither: neither implement nor test the policy.

Experiment 2 ($n = 883$ respondents $\times 3$ vignettes = 2,649 observations): The government always conducted the experiment, and respondents were assigned to one of three conditions revealing the experiment's outcome: 1) Positive Impact: evidence supported the policy; 2) Negative Impact: evidence opposed the policy; 3) No Impact: evidence was inconclusive. The government proceeded with the nationwide rollout only if the results were positive.

In each experiment, respondents evaluated three vignettes (in randomized order) about different policy proposals: 1) an educational program to increase primary school performance; 2) an implicit bias training program for police officers; and 3) an algorithmic decision-making system for welfare payments. Each vignette described a policy problem and a government campaign promise, highlighted potential benefits and costs, and noted researchers' reservations about an immediate rollout. Researchers proposed a randomized trial and recommended proceeding only if the trial demonstrated a positive impact.

2.2.1 Preferences and beliefs about policy impact

Prior to seeing the government decision (Experiment 1) or the result of the randomized policy trial (Experiment 2), we elicited respondent's preferences for experimentation across each of the three policy scenarios. For each vignette, respondents ranked their preferences for the government's action: 1)

[2] Related work on "experiment aversion" in the corporate context (e.g., manipulating the Facebook "News Feeds") has produced mixed conclusions about public support for experimentation (e.g., Meyer et al., 2019; Mislavsky, Dietvorst and Simonsohn, 2020; Mazar, Elbaek and Mitkidis, 2023). A significant limitation of this work is that preferences for experimentation have been elicited without providing an explanation of the logic of experimentation or the meaning of key terms. Public understanding of randomized experiments is relatively low, with recent surveys showing that less than 40% have even heard of the concept (Biddle, Gray and Hiscox, 2023). Without ensuring common knowledge, it is unclear whether stated support for "experimentation" reflects true preferences for experiments or a lack of understanding about the logic of experimentation and/or aversion to terms like "experiment" and "random".

Implement: proceed as planned; 2) Experiment: test before deciding; 3) Neither: do not proceed or test. Respondents also indicated their beliefs about the policy's impact (negative, positive, or no impact) and their confidence in these beliefs on a 5-point scale from "Not at all confident" to "Extremely confident".

After measuring preferences and beliefs, respondents were randomly assigned to a government decision (Experiment 1) or experiment results (Experiment 2). They then evaluated the government's decision on six outcome measures. Additional details about the experiments and treatments are provided in the SI.

2.2.2 Outcome measures

After revealing the government's decision (Experiment 1) or the result of the randomized policy trial (Experiment 2), respondents were presented with six outcome measures in randomized order:

1. *Justifiability*. "The reasons the government provided for its decision are valid."
2. *Arbitrariness*. "The government made its decision after careful consideration."
3. *Revisability*. "The government is open to revising its position in light of new evidence."
4. *Support*. "Do you support or oppose the government's decision?"
5. *Fairness*. "How fair was the government's decision?"
6. *Trust*. "How often do you think this government can be trusted to make the right decisions?"

The first three items, measured using a 7-point scale from"Strongly disagree" to "Strongly agree", gauge perceptions of the legitimacy of government policy decisions along key dimensions identified in recent theoretical work (Tanasoca and Leigh, 2024). The remaining three items capture public support (7-point scale: "Strongly Oppose" to "Strongly Support"), perceived fairness (7-point scale: "Extremely unfair" to "extremely fair"), and broader trust in government decision-making (5-point scale: "Never" to "Always"). All six items tap into the interdependent concepts of normative alignment, fairness, and trust, which have been widely used in prior research to capture public perceptions of government legitimacy (e.g., Levi, Sacks and Tyler, 2009; Peyton, Sierra-Arevalo and Rand, 2019; Peyton, 2020). Responses to all six items were highly correlated ($\alpha = 0.90$), and we constructed a summary index of all outcomes using inverse covariance weighting (Anderson, 2008).

3 Results

Section 3.1 shows descriptive results for the distribution of pre-treatment preferences and beliefs in the full sample. Section 3.2 provides estimates for the average causal effects in each experiment.

3.1 Pre-treatment preferences and beliefs

We find clear evidence that the vast majority of respondents preferred that the government conduct a randomized policy trial before implementing their campaign promise. Overall, 76% of respondents selected experimentation as their first preference, 12% selected implementation, and the remaining 12% preferred that the government abandon their plans without conducting the experiment or proceeding with the implementation. Figure 1 shows the distribution of preferences across each of the three policy scenarios.

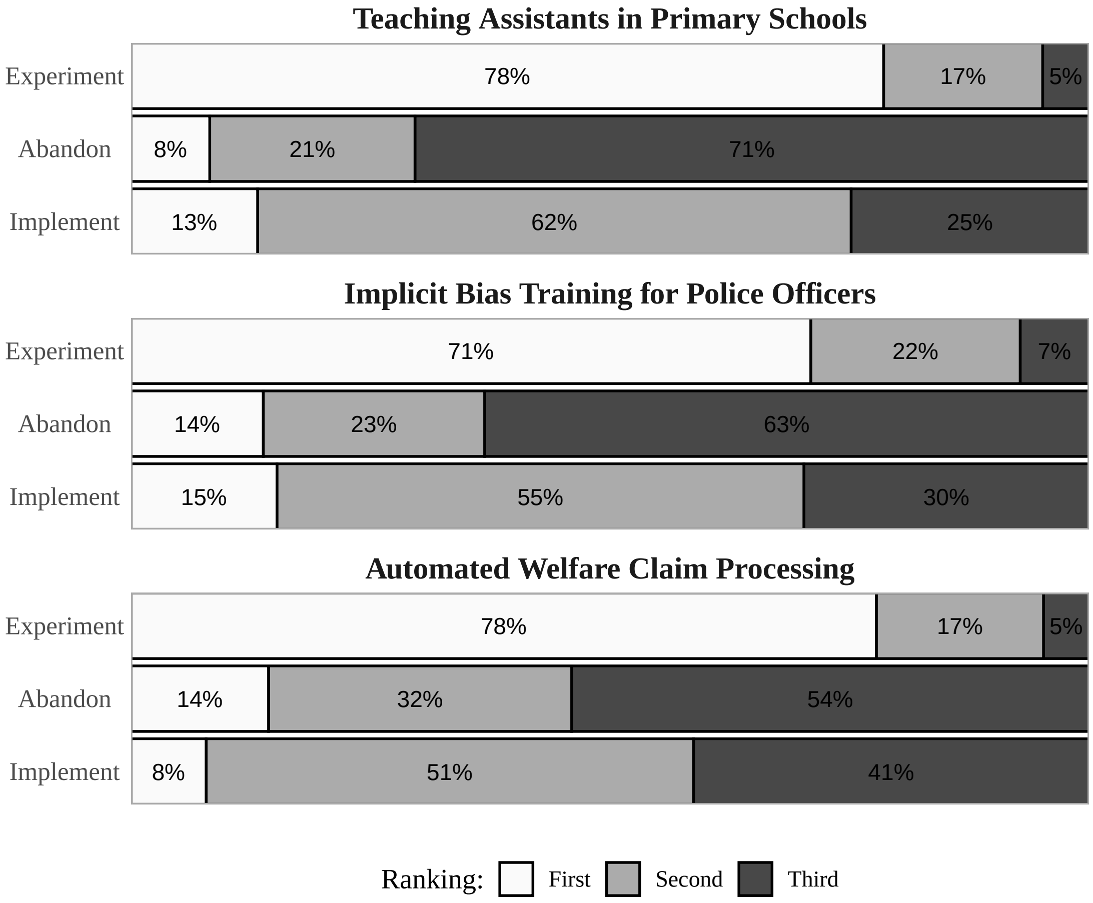


Figure 1: Distribution of preferences over government decisions by policy scenario. Horizontal bars show the distribution of rankings assigned to a given decision for each policy scenario (i.e., 78% gave their first preference to experimentation for the education policy). Proportions are calculated by restricting the sample to the first randomly assigned vignette that respondents evaluated in both experiments (n = 1,782). Labelled percentages may not add exactly to 100% due to rounding.

In every scenario, over 70% of respondents ranked experimentation first, and fewer than 10% ranked it last. We observe some heterogeneity in preferences by policy scenario. For instance, 71% ranked entirely abandoning the campaign promise last for teaching assistants in primary schools, compared to 63% for implicit bias training for police officers and 54% for automated welfare claims processing.

Result 1. *Voters are not averse to experimentation. The vast majority prefer that government conduct randomized experiments to test policy prior to nationwide implementation.*

Figure 2 illustrates how respondents' beliefs about the impact of the government proposal varied across each policy scenario. Approximately 80% believed that increasing the number of teaching assistants in primary schools would positively impact student performance, while only 4% believed it would have a negative impact. Beliefs were more varied for police implicit bias training and automated welfare claim processing. About 50% predicted that implicit bias training would reduce complaints of racial profiling against police, 21% believed it would increase complaints, and 29% thought it would have no impact. Similarly, only 48% believed that an algorithmic decision-making system would increase the accuracy of claims processing, 30% believed it would decrease accuracy, and 22% thought it would have no impact.

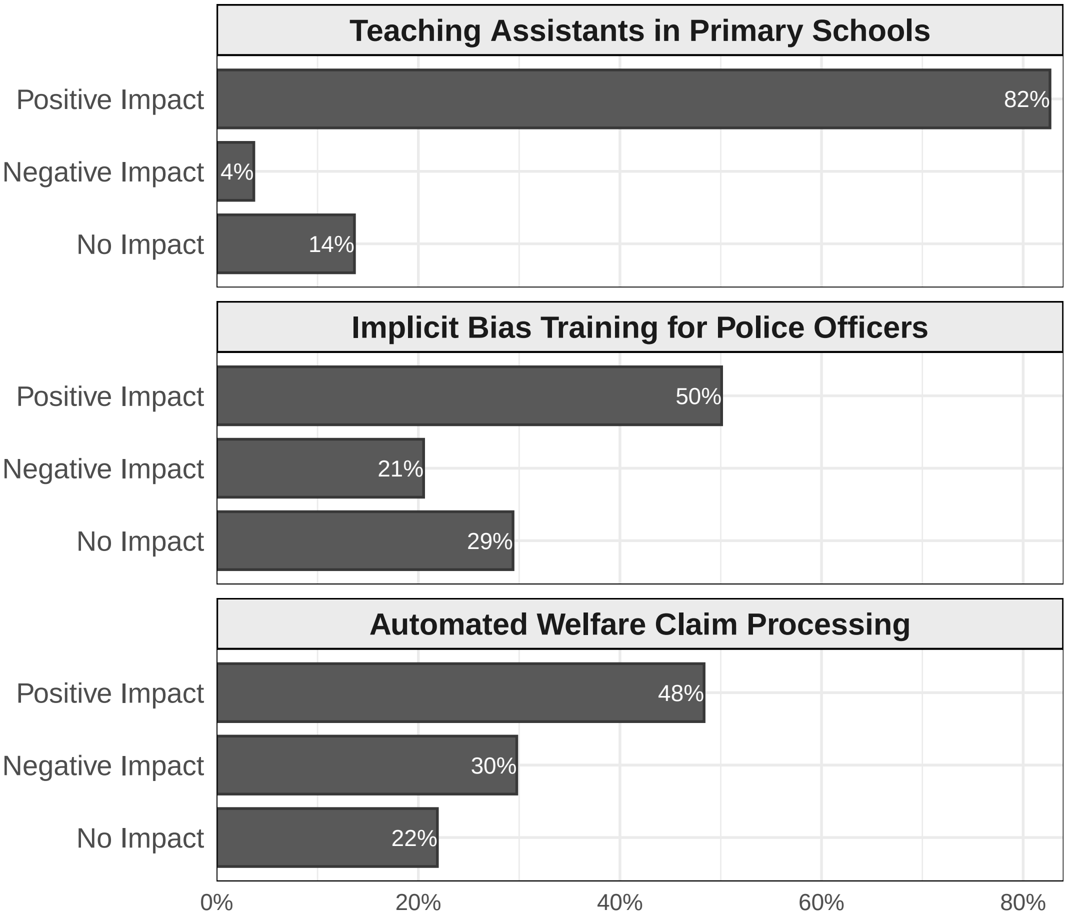


Figure 2: Distribution of beliefs about the impact of government proposal by policy scenario. Proportions are calculated by restricting the sample to the first randomly assigned vignette that respondents evaluated in both experiments (n = 1,782). Labelled percentages may not add exactly to 100% due to rounding.

Despite this variation in beliefs, experimentation was preferred by an absolute majority in all three scenarios. Abandoning the campaign promise was only preferred over implementation when the government proposed an algorithmic decision-making system to automate the processing of Medicare and Centrelink claims. This suggests that preferences for experimentation are not solely based on whether respondents believe a policy will have a positive impact. However, we do find some evidence that preferences for experimentation are correlated with beliefs about the predicted impact of the proposed policy.

Pooling data across all policy scenarios, the probability of selecting experimentation was about 79% among those who believed the policy would have no impact. Among those who believed the policy would have a negative impact, this probability was slightly lower at about 75%, a difference of 4 percentage points (se $= 0.03$, $P = 0.25$). Conversely, the probability of selecting experimentation was significantly higher at 93% among those who believed the policy would have a positive impact, a statistically significant difference of 14 percentage points compared to those who believed it would have no impact (se $= 0.02$, $P < 0.01$).

Result 2. *Preferences for experimentation are not solely determined by beliefs about policy impact. While support for experimentation is higher when a policy is predicted to have a positive impact, an absolute majority prefers experimentation over implementation or abandonment, regardless of whether they believe the policy will have no impact, a positive impact, or a negative impact.*

3.2 Average treatment effects

The primary estimands of interest in Experiment 1 are the average causal effects of a government decision to conduct a randomized policy trial ("Experiment") or do nothing ("Abandon"), relative to a status quo decision to implement policy from their campaign promise without evaluation ("Implement"). In Experiment 2, the primary estimands of interest are the average causal effects of the results of a randomized policy trial ("Positive Impact" or "Negative Impact" versus "No Impact"), given a government decision to conduct a randomized policy trial. As specified in our pre-analysis plan, we use linear regression of the outcome on treatment assignment (reference: "Implement" or "No Impact"), with robust standard errors clustered at the respondent level to correct for within-respondent clustering. To increase precision and reduce the potential for bias due to the ordering of policy scenarios, we adjust for pre-treatment trust in government[3] and the randomized vignette ordering by centering and interacting with treatment assignment (Lin, 2013).

Figure 3 shows the average effects, pooling across all policy scenarios, on perceptions of government legitimacy, as measured by our summary index of all outcome measures. To facilitate interpretation of effect sizes and comparisons, estimates are standardized using Glass's Δ, which scales outcomes by the SD in the reference arm within each experiment (Glass, 1976), and shown with both 90% and 95% confidence intervals (CIs). Dotted vertical lines denote our pre-registered margin of equivalence (MOE) bound of ± 0.20 standard units. This MOE corresponds to one-fifth of 1 SD on the outcome index. When the 90% CI for an estimated effect is contained within the MOE, the null hypothesis of non-equivalence is rejected in favor of equivalence. We conclude that an estimated effect is distinguishable from zero when the 95% CI excludes zero, but "negligible" (i.e., statistically equivalent to ± 0.20

[3] The *Trust in Government Index*, described in Section 2.1.1, was strongly correlated with the outcome index ($\rho = 0.39$).

standard units) when the estimated 90% CI falls within the MOE (Rainey, 2014; Lakens, 2017; Peyton, 2020).

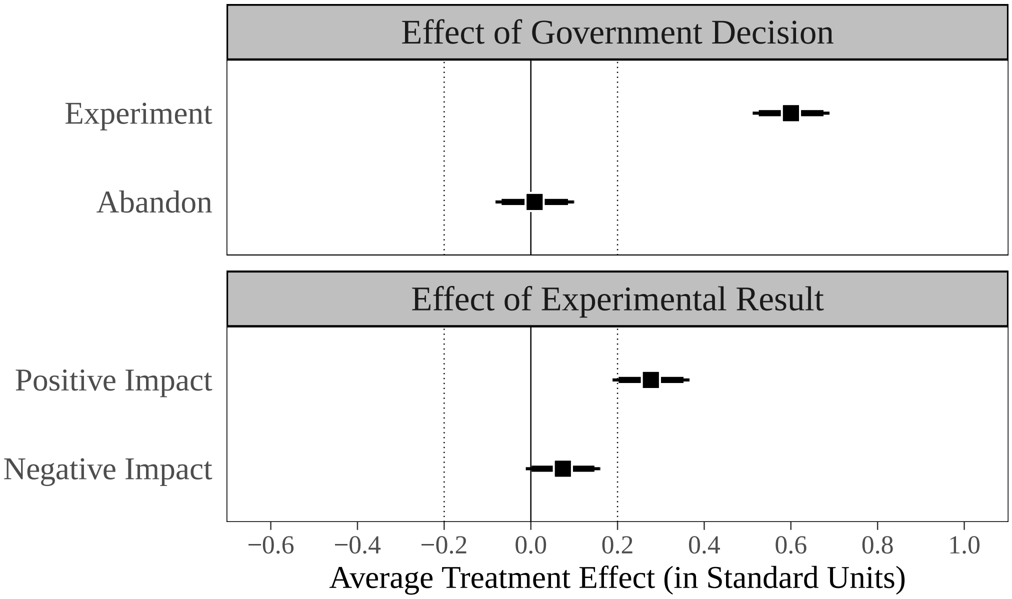


Figure 3: Effects of government decision (Experiment 1, top panel) and experimental results (Experiment 2, bottom panel) on outcome index. Thick horizontal lines denote 90% CIs and thin lines denote 95% CIs. Dotted vertical lines denote an MOE of ± 0.20 standard units. All point estimates (and CIs) from covariate-adjusted linear regression estimator with CIs based on HC2 robust standard errors. To facilitate comparisons, all estimates are standardized using Glass's Δ, which scales outcomes by the standard deviation in the reference arm within each experiment.

Pooling across all policy scenarios, we find that conducting a randomized policy experiment causes a significant increase in public perceptions of government legitimacy ($\delta = 0.60$, se $= 0.04$, $P < 0.01$). This provides clear evidence that randomized policy experiments enhance legitimacy. Moreover, when the effects of a policy are uncertain, abandoning a campaign promise does not necessarily cause a decrease in perceptions of legitimacy relative to proceeding with a planned implementation ($\delta = 0.01$, se $= 0.05$, $P = 0.85$). This negligible effect suggests that implementing untested policies is, on average, comparable to abandoning campaign promises.

We also find that, given a decision to conduct a randomized policy experiment, proceeding with nationwide implementation after a policy has been shown to have a positive impact causes a significant increase in government legitimacy ($\delta = 0.28$, se $= 0.04$, $P < 0.01$). This demonstrates that governments are rewarded for using randomized experiments to inform decisions about policy implementation. Further, abandoning plans for nationwide implementation after a policy has been shown to have a negative impact does not decrease perceptions of government legitimacy ($\delta = 0.07$, se $= 0.04$, $P = 0.09$). This negligible effect shows that abandoning plans for implementation when experiments show a negative impact is comparable to abandoning plans for implementation when experiments show no impact. In other words, governments are not penalized when policy experiments yield negative results.

3.2.1 Heterogeneity by policy scenario in Experiment 1

Figure 4 shows the estimated effects of the government decision when estimated separately for each policy scenario. We find clear evidence that conducting a randomized experiment causes a significant increase in perceptions of government legitimacy across all three policy scenarios. However, the effects of abandoning a campaign promise, relative to proceeding with a planned implementation, vary across policy scenarios.

Conducting an experiment to estimate the effect of teaching assistants in primary schools on student performance, rather than implementing the policy nationwide, causes a significant increase in perceptions of government legitimacy ($\delta = 0.57$, se $= 0.07$, $P < 0.01$). Abandoning this campaign promise, however, causes a significant *decrease* in legitimacy ($\delta = -0.31$, se $= 0.08$, $P < 0.01$).

While conducting an experiment on implicit bias training for police officers also increases legitimacy ($\delta = 0.46$, se $= 0.07$, $P < 0.01$), abandoning this campaign promise is comparable to implementing the policy untested ($\delta = -0.07$, se $= 0.07$, $P = 0.31$). Finally, conducting an experiment on automated welfare claims processing has a significant positive effect on legitimacy ($\delta = 0.78$, se $= 0.07$, $P < 0.01$). Abandoning this campaign promise also causes a significant *increase* in legitimacy ($\delta = 0.42$, se $= 0.07$, $P < 0.01$).

Overall, this demonstrates that conducting randomized experiments consistently increases government legitimacy across heterogeneous policy scenarios. The effects of abandoning campaign promises are, however, generally uncertain. Depending on the policy scenario, abandoning a campaign promise can decrease legitimacy, increase legitimacy, or be comparable to implementing untested policy.

Result 3. *Conducting randomized policy experiments prior to nationwide implementation causes a significant increase in public perceptions of government legitimacy. This finding is consistent across different policy scenarios, demonstrating the generalizability of the positive impact of policy experimentation on government legitimacy. The effect of abandoning a campaign promise, relative to proceeding with a planned implementation, is uncertain and varies across policy scenarios.*

3.2.2 Heterogeneity by policy scenario in Experiment 2

Figure 5 shows how the effects of the experimental results vary across policy scenarios. We find clear evidence that proceeding with nationwide implementation after a policy has been shown to have a positive impact causes a significant increase in government legitimacy across all policy scenarios. Moreover, governments are not penalized for abandoning plans for implementation after a policy has been shown to have a negative impact via a randomized experiment, regardless of the policy scenario.

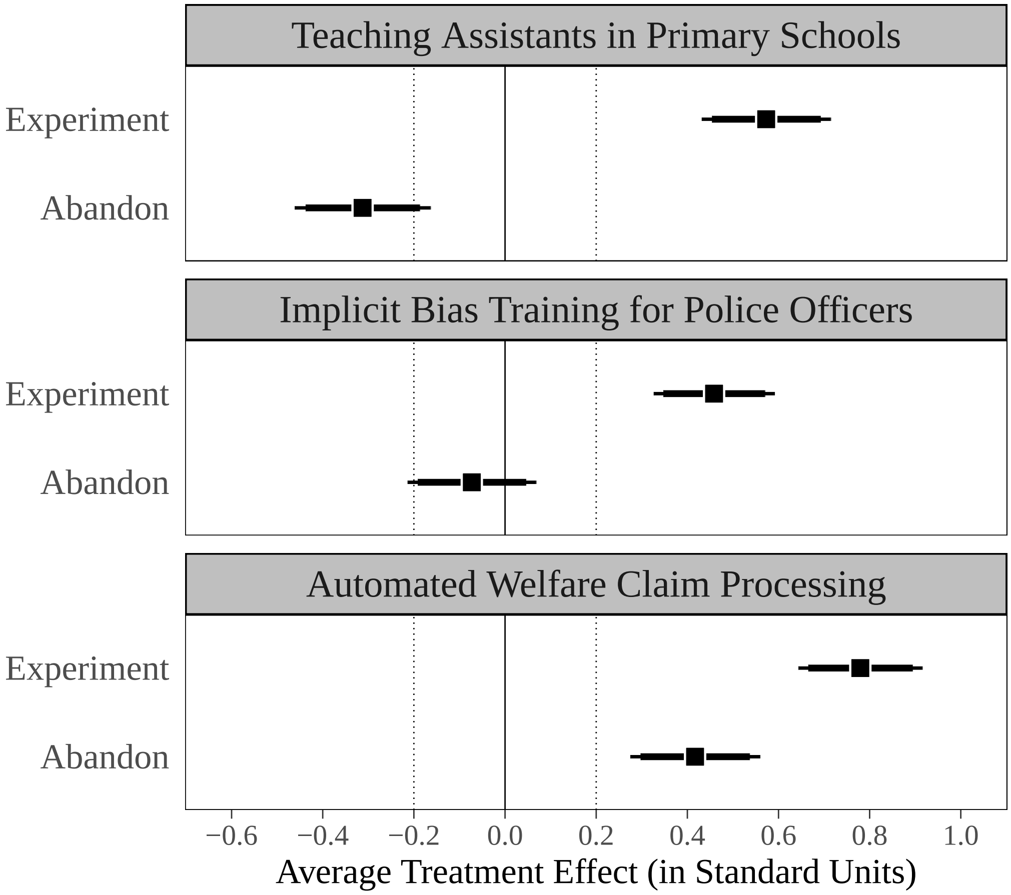


Figure 4: Effects of government decision (Experiment 1) on outcome index by policy scenario. Thick horizontal lines denote 90% CIs and thin lines denote 95% CIs. Dotted vertical lines denote an MOE of ±0.20 standard units. All point estimates (and CIs) are from a covariate-adjusted linear regression estimator with robust standard errors clustered at the respondent level to correct for within-respondent clustering. To facilitate comparisons, all estimates are standardized using Glass's Δ, which scales outcomes by the standard deviation in the reference arm.

Proceeding with nationwide implementation of a policy after an experiment has demonstrated a positive impact has a large positive effect on perceptions of legitimacy when the policy involves increasing the number of teaching assistants in primary schools ($\delta = 0.46$, se $= 0.08$, $P < 0.01$). Abandoning plans for nationwide implementation of this policy after an experiment yields a negative result has a negligible effect on legitimacy that is indistinguishable from zero ($\delta = -0.05$, se $= 0.08$, $P = 0.58$). A similar result holds for the policy mandating implicit bias training for police officers (Positive Impact: $\delta = 0.22$, se $= 0.07$, $P < 0.01$; Negative Impact: $\delta = -0.05$, se $= 0.07$, $P = 0.49$).

Implementing an algorithmic decision-making system for automated processing of welfare claims after an experiment demonstrates a positive impact has a positive effect on perceptions of legitimacy ($\delta = 0.34$, se $= 0.07$, $P < 0.01$). When the results of this experiment are negative and the government

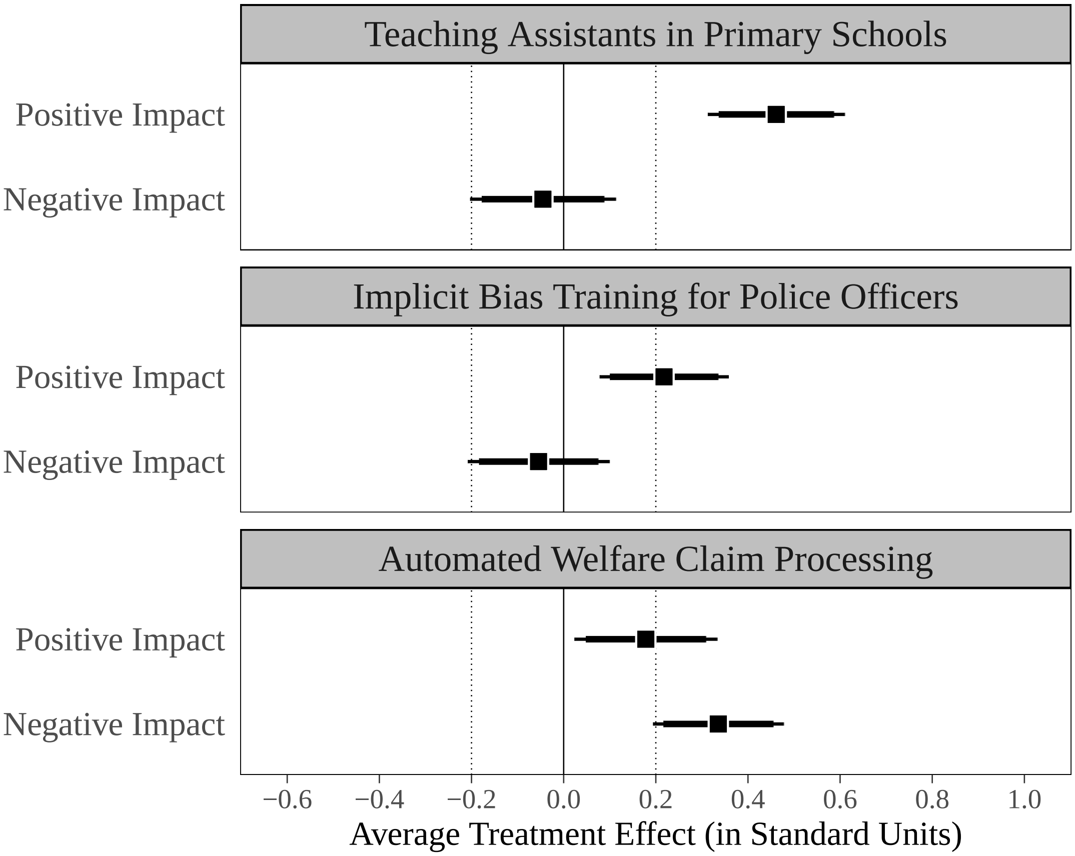


Figure 5: Effects of experimental results (Experiment 2) on outcome index by policy scenario. Thick horizontal lines denote 90% CIs and thin lines denote 95% CIs. Dotted vertical lines denote an MOE of ± 0.20 standard units. All point estimates (and CIs) are from a covariate-adjusted linear regression estimator with robust standard errors clustered at the respondent level to correct for within-respondent clustering. To facilitate comparisons, all estimates are standardized using Glass's Δ, which scales outcomes by the standard deviation in the reference arm.

decides to abandon plans for implementation, this also has a positive effect on legitimacy ($\delta = 0.18$, se $= 0.08$, $P = 0.02$). This demonstrates that governments are rewarded when policy implementation is guided by positive experimental results, and not penalized when implementation plans are abandoned after a policy is shown to have a negative impact. These findings generalize across policy scenarios.

Result 4. *Government legitimacy is maintained when policy decisions are based on experimental evidence, regardless of the results. Implementing a policy after positive experimental results increases legitimacy, while abandoning plans after negative results does not decrease legitimacy. This finding is consistent across different policy scenarios, even if it means retracting planned policies.*

3.2.3 Robustness to public understanding of randomized experiments

This section examines the robustness of our results to the level of public understanding of randomized experiments. Given that public understanding of randomized experiments is relatively low (see e.g., Kennedy and Hefferon, 2019; Biddle, Gray and Hiscox, 2023), ensuring that respondents grasp the key concepts is crucial for interpreting their preferences accurately. To address this, we provided an educational primer on randomized controlled trials (RCTs) and assessed comprehension with three questions covering the concepts of random assignment, the primary purpose of experimentation, and inference from group comparisons (see SI Appendix Fig. S8).

Of the 1,781 respondents, 1,066 (60%) answered all three comprehension questions correctly after reading the educational primer. The remaining 715 respondents were shown the questions they answered incorrectly and provided with the correct answers before proceeding to the next stage of the survey. We partition the sample into two sub-groups: those who demonstrated an understanding by answering all three questions correctly ("understood") and those who answered one or more questions incorrectly ("misunderstood").

Respondents who demonstrated an understanding were more likely to select experimentation as their first preference for what the government should do on the first randomly assigned policy scenario they were presented with. On average, 84% of those who understood reported a first preference for experimentation, whereas only 64% of those who misunderstood selected experimentation as their first preference – a difference of 20 percentage points.

This significant difference suggests that comprehension of the principles and benefits of RCTs influences preferences for policy experimentation. When respondents understand the logic behind randomized trials, they are more likely to recognize the value of policy decisions based on experimental evidence and therefore show greater support for experimentation.

Table 1 shows the estimated treatment effects for each of these subgroups and their differences. In Experiment 1, we find positive treatment effects for the government decision to experiment among both sub-groups; however, there is strong evidence of treatment effect heterogeneity. The average effect among those that demonstrated understanding is significantly higher than among those that did not.

One possible explanation is that those who demonstrated an understanding of RCTs perceive the government's decision to experiment prior to a nationwide rollout as a more informed and legitimate approach to policy implementation. This could enhance their perception of government legitimacy more strongly. However, it is important to note that both groups still show positive treatment effects, suggesting that even those with less understanding of RCTs recognize some value in experimentation.

We find positive treatment effects for the government decision to abandon plans for policy implementation among both sub-groups, indicating that comprehension of RCTs does not influence perceptions of legitimacy when the government decides to abandon a policy. This suggests that abandoning a policy is viewed similarly by both groups, possibly as a failure to fulfill a promise, regardless of the rationale provided.

In Experiment 2, where the government decided to conduct the RCT and the results were randomly assigned (Positive Impact vs. No Impact; Negative vs. No Impact), we do not find evidence of differences in treatment effects between those who demonstrated understanding and those who did not. This suggests that once the decision to conduct a policy experiment has been made, the understanding of the basic logic of RCTs may be less important. Both groups, regardless of their initial comprehension levels, may equally understand the implications of positive and negative results.

	Understood	**Misunderstood**	**Difference**
Effect of Government Decision			
Experiment	0.86 (0.06)*	0.24 (0.06)*	0.62 (0.09)*
Abandon	0.03 (0.06)	-0.02 (0.06)	0.05 (0.09)
Effect of Experimental Result			
Positive Impact	0.28 (0.06)*	0.28 (0.07)*	-0.00 (0.09)
Negative Impact	0.09 (0.05)	0.06 (0.07)	0.03 (0.09)

Table 1: Effects of government decision (Experiment 1) and experimental result (Experiment 2) on outcome index by demonstrated understanding of randomized experiments. $^*P < 0.05$.

Result 5. *Public understanding strongly influences preferences for experimentation and the effects that randomized policy trials have on legitimacy. Respondents who demonstrated an understanding of experimentation were significantly more likely to prefer experimentation. Government decisions to conduct randomized policy experiments also had larger effects on legitimacy among these individuals.*

4 Conclusion

The findings reported here have both practical and theoretical implications for public policy and governance. In addition to the potential benefits that randomized policy experiments hold for improving human welfare at scale, they may also increase government legitimacy by mitigating normative concerns raised by the implementation of untested policies without credible evidence of their potential benefits or harms (Tanasoca and Leigh, 2024). Despite these potential benefits, recent empirical work suggests that public preferences for universal implementation over experimentation pose a significant barrier to evidence-based policy (e.g., Meyer et al. 2019; Heck et al. 2020; but see Mislavsky, Dietvorst and Simonsohn 2020; Mazar, Elbaek and Mitkidis 2023)

To date, however, little is known about public support for randomized policy experiments in the context of government decision-making. Our results shed new light on this issue, showing that the public overwhelmingly supports randomized policy experiments over universal implementation of untested government policy. This preference spans various policy domains, indicating that voters value government decisions guided by experimental evidence.

More importantly, we found that decisions to conduct randomized policy experiments lead to a substantial increase in public perceptions of government legitimacy. This effect is consistent across different policy domains, demonstrating the broad positive impact of policy experimentation on government legitimacy. Furthermore, we found that government legitimacy remains intact when policy implementation is based on experimental evidence, regardless of the results. Implementing a policy after an experiment demonstrates a positive impact increases legitimacy, but abandoning plans for implementation if an experiment demonstrates a negative impact does not decrease legitimacy.

Public understanding of randomized experiments plays a crucial role in these dynamics. Individuals who understood the logic of experimentation expressed even stronger preferences for randomized policy experiments. While government decisions to conduct randomized policy experiments positively affected legitimacy overall, these effects were even greater among those who demonstrated an understanding of experimentation. This highlights the important role that public education can play in building support for evidence-based policy.

In conclusion, our research suggests that governments would benefit from increasing their focus on randomized policy experiments. Conducting these experiments not only enhances policy effectiveness but also increases public trust and the perceived legitimacy of government decision-making. By embracing a scientific approach to policy implementation, governments can justify their actions with credible evidence and mitigate normative concerns raised by the status quo practice of implementing untested policies on large populations without credible evidence of their potential benefits or harms.

Supporting Information Appendix

Following best practices to ensure data quality in online survey sampling (Berinsky, Margolis and Sances, 2014; Peyton, Huber and Coppock, 2021; Ternovski and Orr, 2022), we restricted participation to respondents that passed an attention screener placed at the beginning of the survey. Screening out respondents based on attention checks placed near the end of the survey or after treatments are administered can induce bias, but using attention checks administered early in the survey or prior to experimental treatments does not induce bias (Montgomery, Nyhan and Torres, 2018; Aronow, Baron and Pinson, 2019). We employed an attention screener used in recent survey experimental work (Peyton, Huber and Coppock, 2021; Vaughn, Peyton and Huber, 2022; Peyton, Weiss and Vaughn, 2022). After viewing the screener, participants were asked two attention check questions shown in Fig. S7 with correct responses are highlighted in bold text.

Among the 2,993 individuals that consented to participate (see Fig. S6), 60% passed the first attention check question (ACQ) and went on to complete our survey. Among these 1,781 individuals, the median time to completion was 12 minutes, and 64% ($n = 1,143$) also passed the second ACQ. Only those individuals that provided incorrect answers to the first ACQ, or refused to answer, were terminated from the survey. Those that passed the ACQ next responded to a set of questions covering demographics (e.g., age, education) and political attitudes (e.g., trust in government and partisanship).

Prior to treatment assignment, all respondents read a short educational primer explaining the logic of randomized experiments and the meaning of key concepts like "random assignment". Next, they responded to three comprehension questions covering the concept of random assignment, the purpose of experimentation, and inference from group comparisons (see Fig. S8 with correct responses highlighted in bold). 1,066 respondents (60%) answered all three questions correctly and proceeded immediately to the next stage. The remaining 715 respondents were first shown the questions they answered incorrectly and provided with the correct answer before moving to the next stage (i.e., none were excluded for answering incorrectly).

All 1,781 respondents were randomly assigned to one of two parallel experiments with equal probability using simple random assignment. In each experiment, respondents were shown three vignettes (in randomized order) about three different policy proposals: 1) an educational program to increase student performance in primary schools (see Fig. S9); 2) an implicit bias training program to decrease racial profiling by police officers (see Fig. S10); and 3) an algorithmic decision-making system for welfare payments to improve the accuracy of claims processing (see Fig. S11).

The vignettes presented to respondents were identical aside from the content specific to each policy domain. Each described a policy problem (e.g., declining student performance) and a campaign promise made by a recently elected government to address the problem (e.g., increasing the number of teaching assistants in primary schools). Each vignette also highlighted the potential benefits and costs of implementing the policy, and noted that researchers had expressed reservations about an immediate nationwide rollout. The researchers had prepared a submission to the government proposing a randomized trial to first evaluate the government's policy proposal, and a recommendation to only proceed with a nationwide rollout if the results demonstrated that the policy had the intended effects.

For each vignette, respondents were first asked to rank order their preferences for what the government should do from three options: 1) **Implement**: proceed with the nationwide rollout as initially planned; 2) **Experiment**: test the program with a randomized trial before deciding on the nationwide rollout; 3) **Neither**: do not proceed with the nationwide rollout and do not test the program with a randomized trial. Respondents' beliefs about the effect of the policy proposal were then elicited. Specifically, they were asked whether they believed it would have: 1) a **Negative impact** (e.g., "decrease student performance"); 2) a **Positive impact** (e.g., "increase student performance"); or 3) **No impact** (e.g., "no impact on student performance"). They were also asked how confident they were about the answer they provided using a 5-point scale from "Not at all confident" to "Extremely confident."

After measuring respondents' preferences and beliefs, they were randomly assigned either a government decision (Experiment 1, $n = 898$ respondents) or the results of a randomized experiment evaluating the government's policy proposal (Experiment 2, $n = 883$ respondents). Table S2 shows the distribution of treatment assignments by policy scenario for each experiment.

In Experiment 1 ($n = 898$ respondents ×3 vignettes = 2,694 observations), respondents were independently assigned one of three conditions with equal probability for each policy scenario: 1) **Implement**: the government decided to implement the new policy as planned without conducting the experiment; 2) **Experiment**: the government decided to first conduct the randomized experiment before deciding on whether to implement the policy; or 3) **Neither**: the government decided to neither implement the policy nor conduct the experiment. Figures S12-S14 show each treatment in the education vignette.

In Experiment 2 ($n = 883$ respondents ×3 vignettes = 2,649 observations), the government always decided to conduct the experiment (i.e., the "Experiment" condition in Experiment 1). However, respondents were independently assigned to one of three conditions revealing the result of the experiment for each policy scenario: 1) **Positive Impact**: the experiment provided evidence in support of the government's policy; 2) **Negative Impact**: the experiment provided evidence against the government's policy; or 3) **No Impact**: the experiment provided inconclusive evidence. The government proceeded with the nationwide rollout only when the results were positive; otherwise, they abandoned plans for the nationwide rollout. Figures S15-S17 show each treatment in the education vignette.

	Experiment 1			**Experiment 2**		
	Implement	Abandon	Experiment	No Impact	Positive Impact	Negative Impact
Teaching Assistants in Primary Schools	293	299	306	289	304	290
Implicit Bias Training for Police Officers	303	286	309	293	289	301
Automated Welfare Claim Processing	300	297	301	298	302	283
Total:	896	882	916	880	895	874

Table S2: Distribution of treatment assignments by policy scenario. Experiment 1: 898 respondents evaluated 3 policy scenarios in randomized order (2,694 observations). Experiment 2: 883 respondents evaluated 3 policy scenarios in randomized order (2,649 observations). Within each experiment, treatments were independently assigned to each policy scenario with equal probability using simple random assignment.

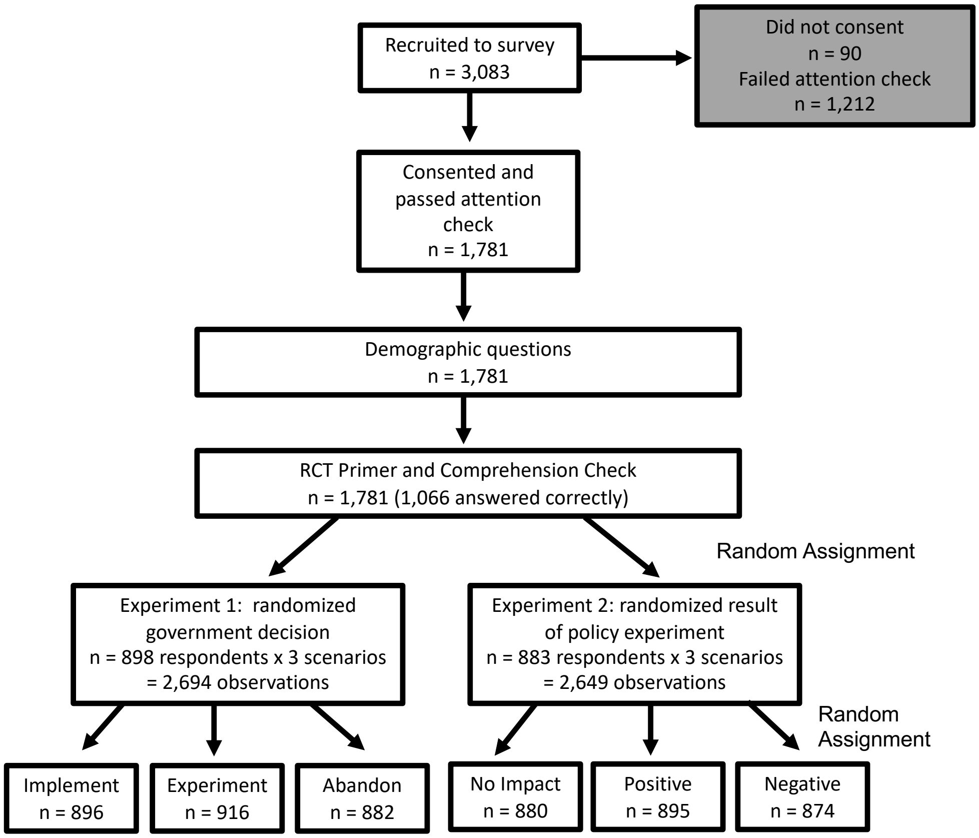
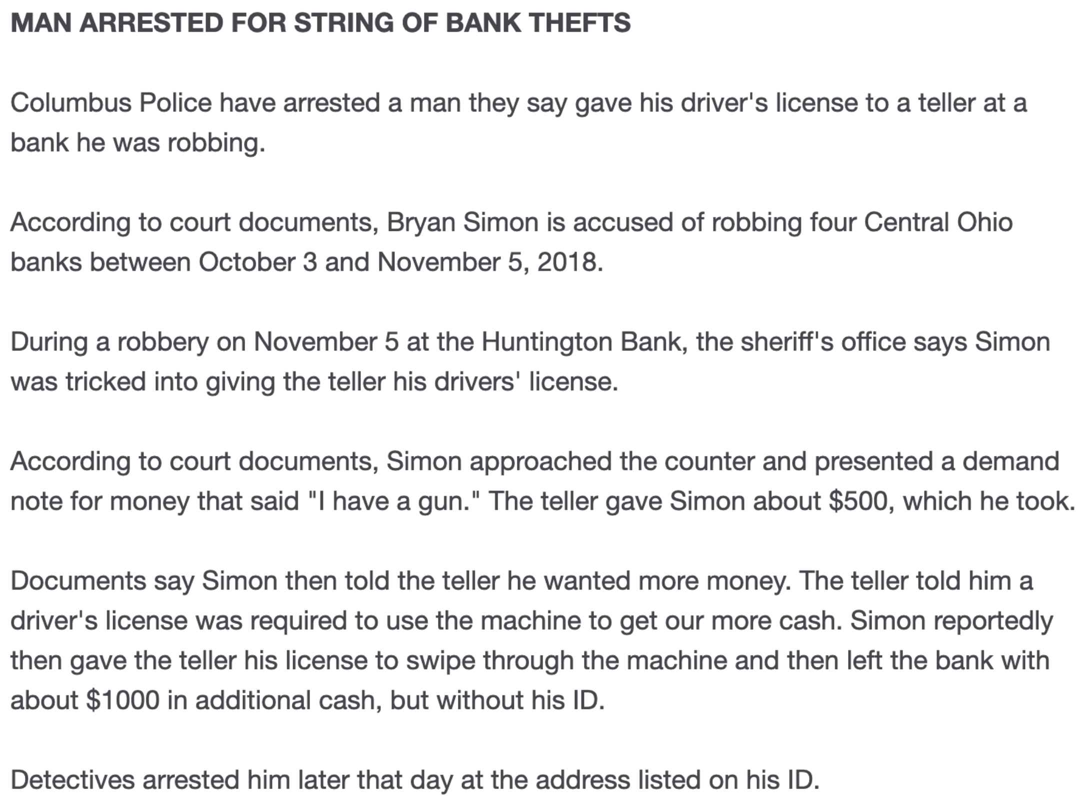


Figure S6: Summary of experimental design and randomizations.

Variable (level)		Sample Proportion	Target Proportion	Difference
Female:				
	Yes	0.49	0.51	-0.02
	No	0.51	0.49	0.02
State/Territory:				
	New South Wales	0.32	0.32	0
	Victoria	0.27	0.26	0.01
	Queensland	0.20	0.20	0
	Western Australia	0.10	0.10	0
	South Australia	0.08	0.07	0.01
	Tasmania	0.02	0.02	0
	Australian Capital Territory	0.02	0.02	0
	Northern Territory	< 0.01	0.01	< 0.01
Age:				
	18-24	0.12	0.11	0.01
	25-34	0.17	0.18	-0.01
	35-44	0.19	0.18	0.01
	45-54	0.16	0.16	0
	55-64	0.15	0.15	0
	65+	0.20	0.22	-0.02
Education:				
	Less than year 12	0.10	-	-
	Year 12	0.24	-	-
	Vocational	0.25	-	-
	Incomplete University	0.08	-	-
	Bachelor degree	0.25	-	-
	Postgraduate degree	0.08	-	-
Household income:				
	$29,999 or less	0.16	-	-
	$30,000 - $49,999	0.17	-	-
	$50,000 - $79,999	0.22	-	-
	$80,000 - $99,999	0.13	-	-
	$100,000 - $149,999	0.17	-	-
	$150,000 - $199,999	0.09	-	-
	$200,000 and above	0.06	-	-

Table S3: Demographic characteristics for the sample and Australian population. The target proportions for the Australian population come the the Australian Bureau of Statistics. Comparable target proportions are not available for education or household income. In 2023, 32% of persons aged 15-74 held a bachelor degree or above in Australia. In our sample, 33% held a bachelor degree or above. In 2022, median household income in Australia was $92,856 per year. In our sample, 68% reported household income of $99,999 or less.



MAN ARRESTED FOR STRING OF BANK THEFTS

Columbus Police have arrested a man they say gave his driver's license to a teller at a bank he was robbing.

According to court documents, Bryan Simon is accused of robbing four Central Ohio banks between October 3 and November 5, 2018.

During a robbery on November 5 at the Huntington Bank, the sheriff's office says Simon was tricked into giving the teller his drivers' license.

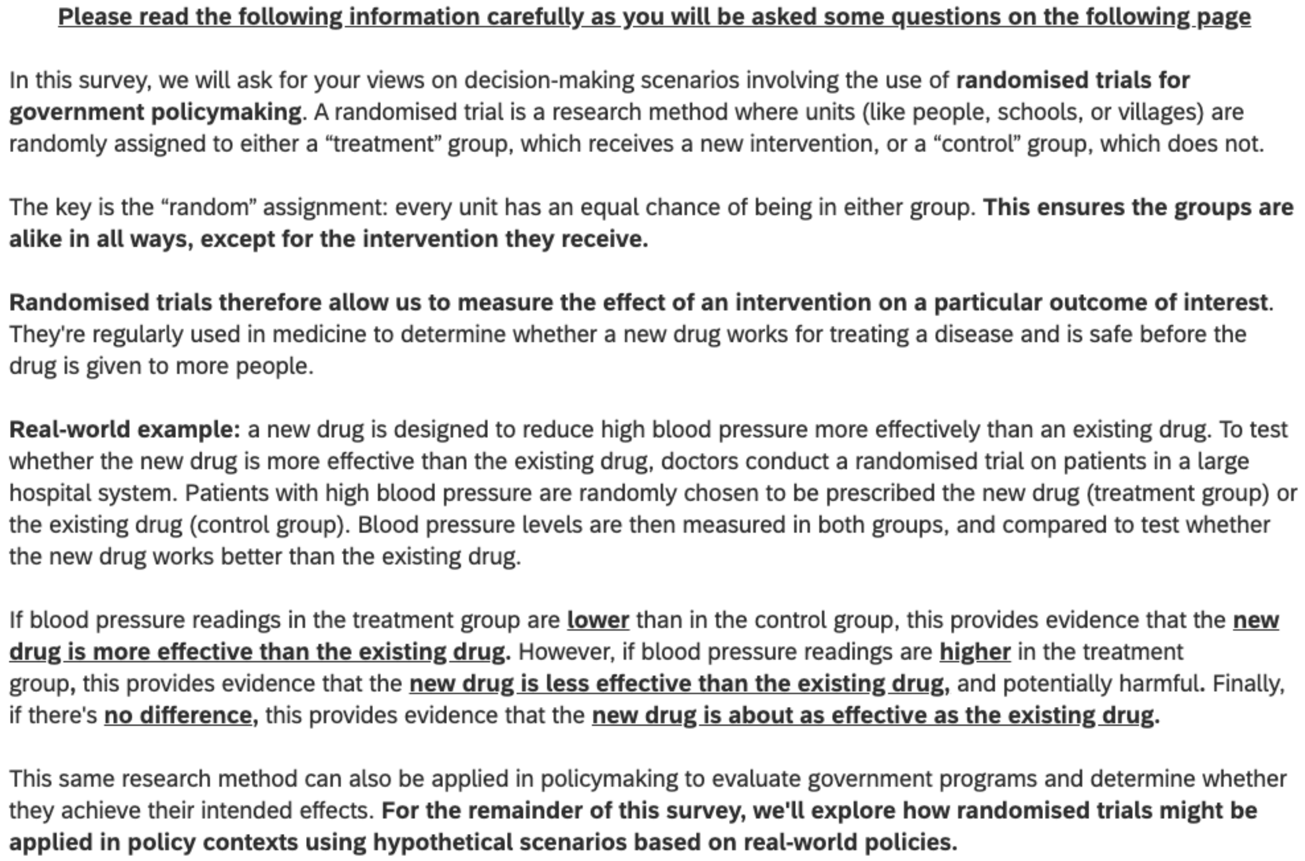
According to court documents, Simon approached the counter and presented a demand note for money that said "I have a gun." The teller gave Simon about $500, which he took.

Documents say Simon then told the teller he wanted more money. The teller told him a driver's license was required to use the machine to get our more cash. Simon reportedly then gave the teller his license to swipe through the machine and then left the bank with about $1000 in additional cash, but without his ID.

Detectives arrested him later that day at the address listed on his ID.

1. How was Simon identified by police for the crime he allegedly committed? [A police officer recognized him, From video surveillance, **Because he left his ID**, He turned himself in, None of the above]
2. How much money did Simon allegedly steal? [About $500, **About $1500**, About $25,000, About $1 million dollars, None of the above]

Figure S7: Attention screener



Please read the following information carefully as you will be asked some questions on the following page

In this survey, we will ask for your views on decision-making scenarios involving the use of **randomised trials for government policymaking**. A randomised trial is a research method where units (like people, schools, or villages) are randomly assigned to either a “treatment” group, which receives a new intervention, or a “control” group, which does not.

The key is the “random” assignment: every unit has an equal chance of being in either group. **This ensures the groups are alike in all ways, except for the intervention they receive.**

Randomised trials therefore allow us to measure the effect of an intervention on a particular outcome of interest. They're regularly used in medicine to determine whether a new drug works for treating a disease and is safe before the drug is given to more people.

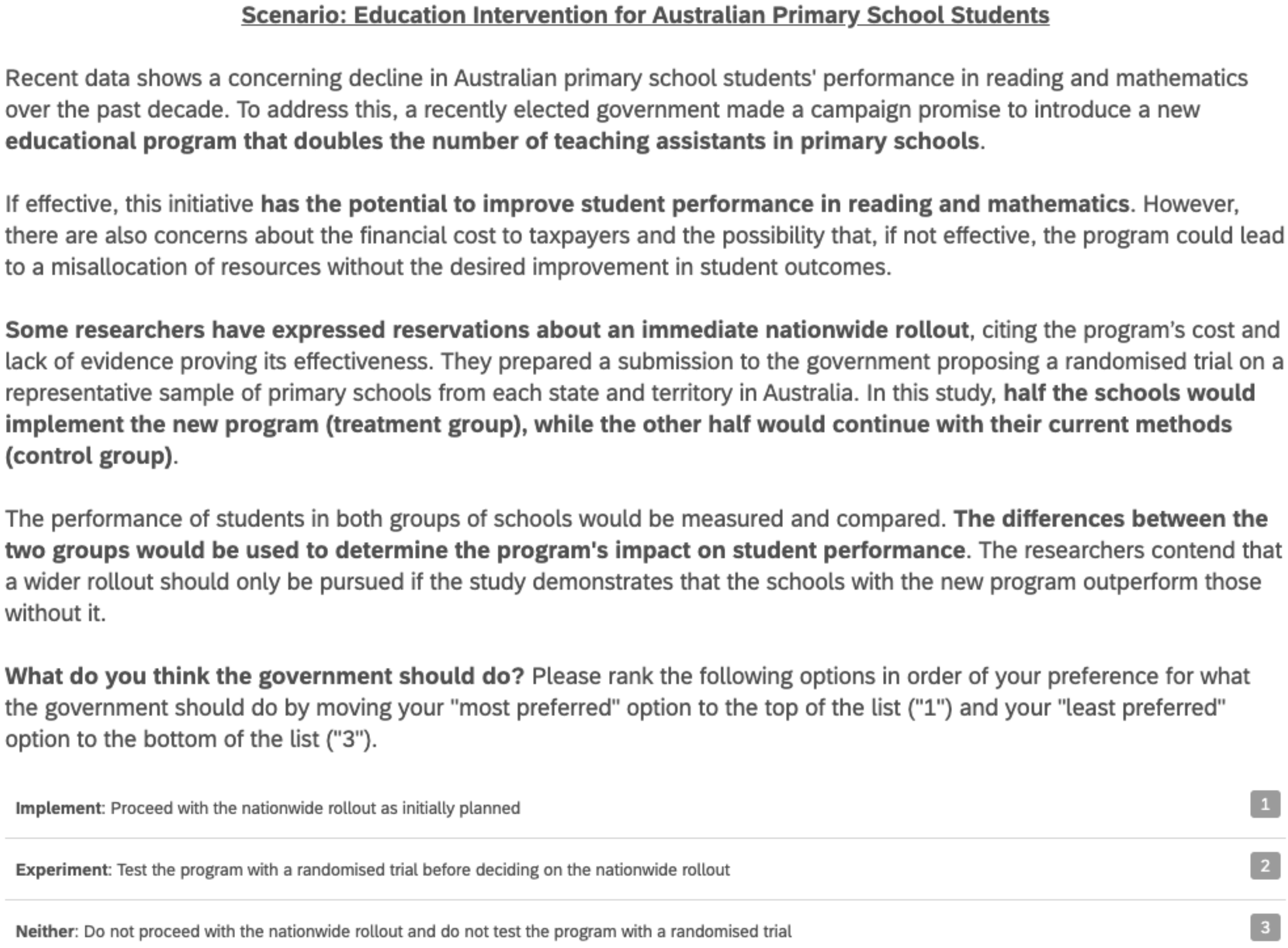
Real-world example: a new drug is designed to reduce high blood pressure more effectively than an existing drug. To test whether the new drug is more effective than the existing drug, doctors conduct a randomised trial on patients in a large hospital system. Patients with high blood pressure are randomly chosen to be prescribed the new drug (treatment group) or the existing drug (control group). Blood pressure levels are then measured in both groups, and compared to test whether the new drug works better than the existing drug.

If blood pressure readings in the treatment group are **lower** than in the control group, this provides evidence that the **new drug is more effective than the existing drug.** However, if blood pressure readings are **higher** in the treatment group**,** this provides evidence that the **new drug is less effective than the existing drug,** and potentially harmful**.** Finally, if there's **no difference,** this provides evidence that the **new drug is about as effective as the existing drug.**

This same research method can also be applied in policymaking to evaluate government programs and determine whether they achieve their intended effects. **For the remainder of this survey, we'll explore how randomised trials might be applied in policy contexts using hypothetical scenarios based on real-world policies.**

1. Why are units (like people or schools) randomly assigned to either the treatment or control group in a randomised trial? [**To ensure both groups are similar, except for the intervention they receive**; To decide which group will receive the more expensive treatment; To give preference to certain units over others; To make the study more complicated]
2. What is the primary purpose of conducting a randomised trial in the context of a new intervention or policy? [To market the intervention to a wider audience; **To determine whether the intervention causes a particular outcome to happen**; To ensure that every unit receives the intervention; To randomly select participants for a study]
3. How do researchers typically determine the effectiveness of an intervention with a randomised trial? [**By comparing the outcomes of the group that received the intervention (treatment group) with the group that did not (control group)**; By assessing the opinions of those who designed the intervention; By checking if the intervention has been adopted in other studies; By asking the group that receive the intervention (treatment group) how they feel after receiving the intervention]

Figure S8: Educational primer on randomized experiments



Scenario: Education Intervention for Australian Primary School Students

Recent data shows a concerning decline in Australian primary school students' performance in reading and mathematics over the past decade. To address this, a recently elected government made a campaign promise to introduce a new **educational program that doubles the number of teaching assistants in primary schools**.

If effective, this initiative **has the potential to improve student performance in reading and mathematics**. However, there are also concerns about the financial cost to taxpayers and the possibility that, if not effective, the program could lead to a misallocation of resources without the desired improvement in student outcomes.

Some researchers have expressed reservations about an immediate nationwide rollout, citing the program's cost and lack of evidence proving its effectiveness. They prepared a submission to the government proposing a randomised trial on a representative sample of primary schools from each state and territory in Australia. In this study, **half the schools would implement the new program (treatment group), while the other half would continue with their current methods (control group)**.

The performance of students in both groups of schools would be measured and compared. **The differences between the two groups would be used to determine the program's impact on student performance**. The researchers contend that a wider rollout should only be pursued if the study demonstrates that the schools with the new program outperform those without it.

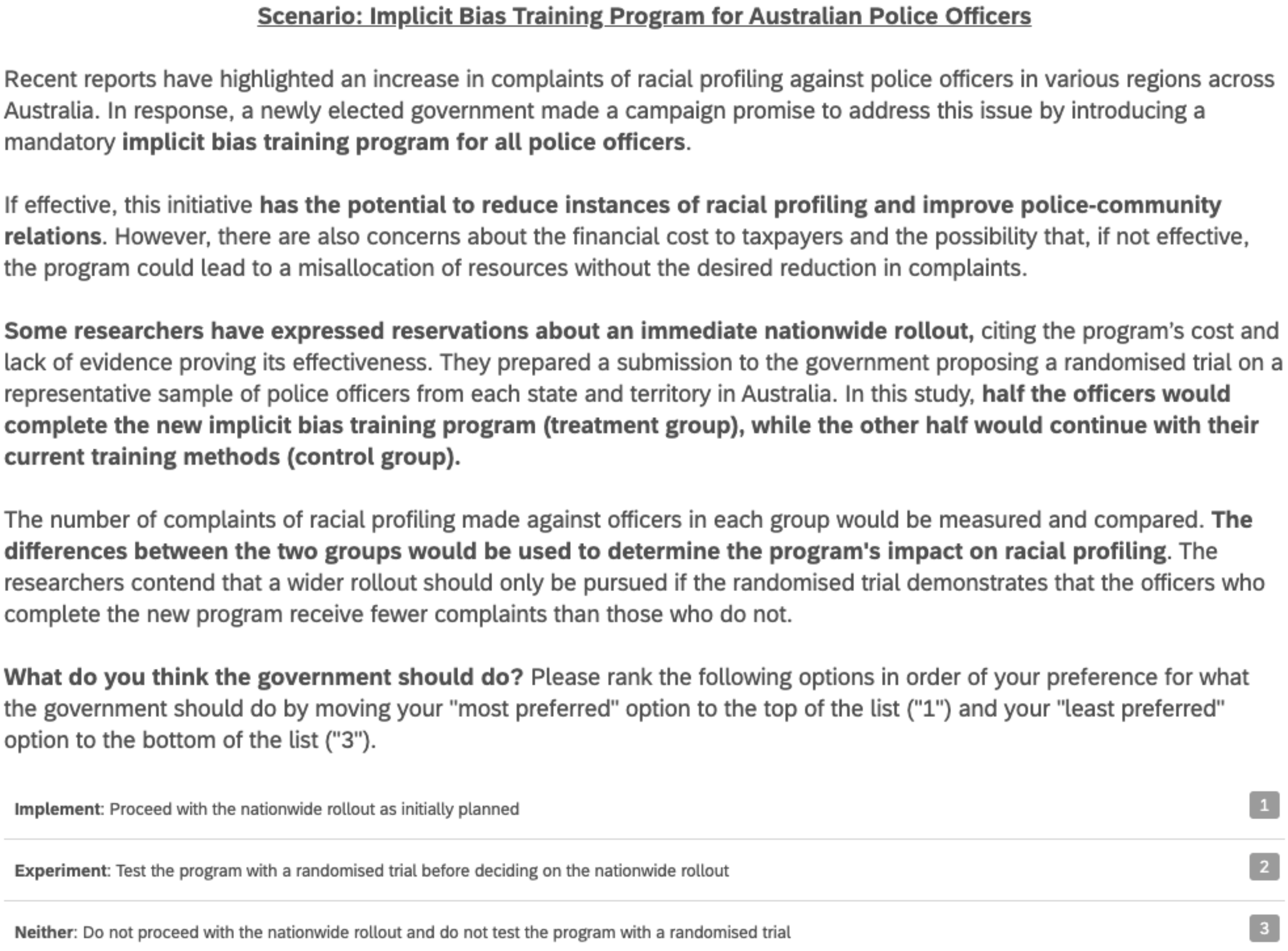
What do you think the government should do? Please rank the following options in order of your preference for what the government should do by moving your "most preferred" option to the top of the list ("1") and your "least preferred" option to the bottom of the list ("3").

Implement: Proceed with the nationwide rollout as initially planned 1

Experiment: Test the program with a randomised trial before deciding on the nationwide rollout 2

Neither: Do not proceed with the nationwide rollout and do not test the program with a randomised trial 3

Figure S9: Teaching assistants in primary schools



Scenario: Implicit Bias Training Program for Australian Police Officers

Recent reports have highlighted an increase in complaints of racial profiling against police officers in various regions across Australia. In response, a newly elected government made a campaign promise to address this issue by introducing a mandatory **implicit bias training program for all police officers**.

If effective, this initiative **has the potential to reduce instances of racial profiling and improve police-community relations**. However, there are also concerns about the financial cost to taxpayers and the possibility that, if not effective, the program could lead to a misallocation of resources without the desired reduction in complaints.

Some researchers have expressed reservations about an immediate nationwide rollout, citing the program's cost and lack of evidence proving its effectiveness. They prepared a submission to the government proposing a randomised trial on a representative sample of police officers from each state and territory in Australia. In this study, **half the officers would complete the new implicit bias training program (treatment group), while the other half would continue with their current training methods (control group).**

The number of complaints of racial profiling made against officers in each group would be measured and compared. **The differences between the two groups would be used to determine the program's impact on racial profiling**. The researchers contend that a wider rollout should only be pursued if the randomised trial demonstrates that the officers who complete the new program receive fewer complaints than those who do not.

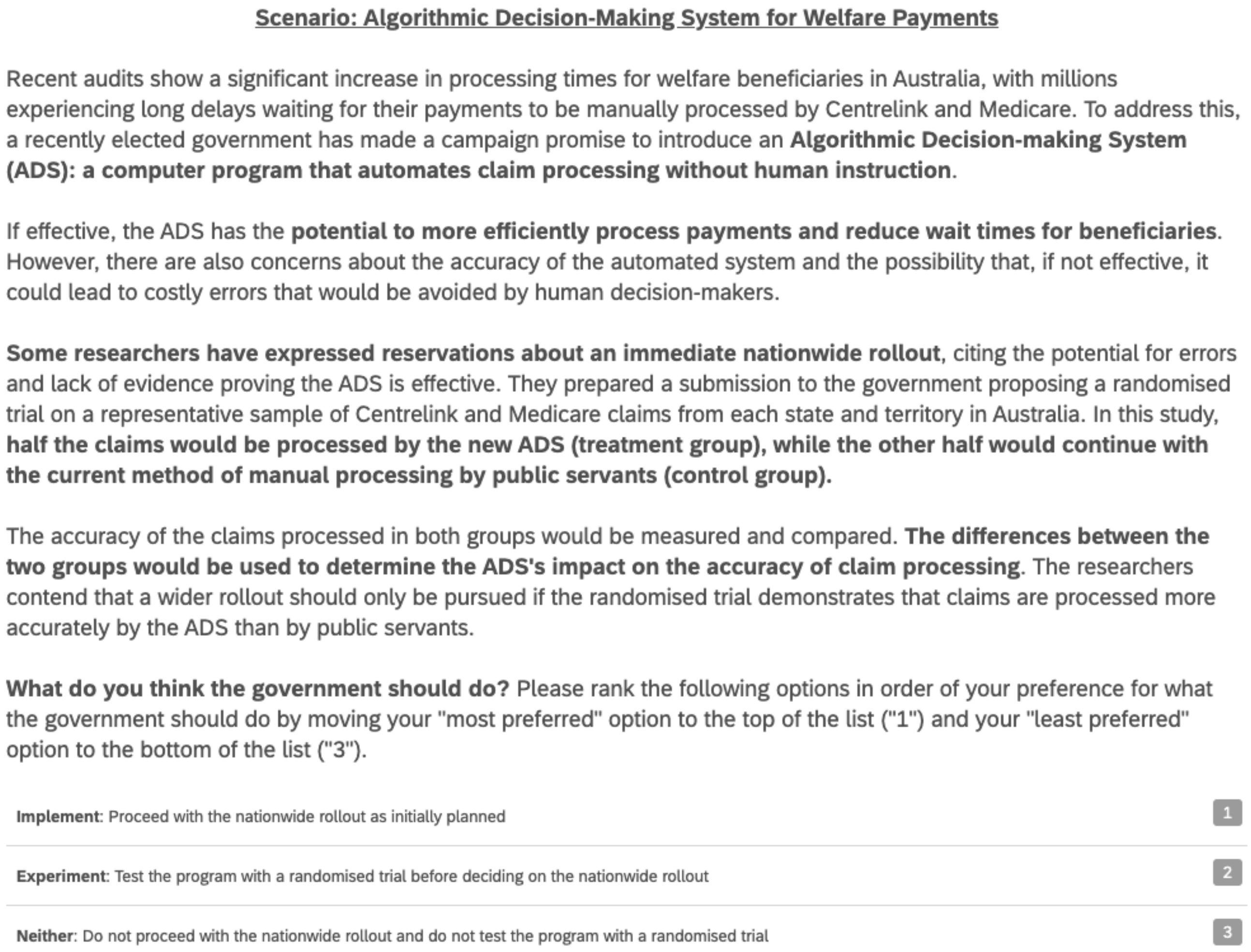
What do you think the government should do? Please rank the following options in order of your preference for what the government should do by moving your "most preferred" option to the top of the list ("1") and your "least preferred" option to the bottom of the list ("3").

Implement: Proceed with the nationwide rollout as initially planned 1

Experiment: Test the program with a randomised trial before deciding on the nationwide rollout 2

Neither: Do not proceed with the nationwide rollout and do not test the program with a randomised trial 3

Figure S10: Implicit bias training for police officers



Scenario: Algorithmic Decision-Making System for Welfare Payments

Recent audits show a significant increase in processing times for welfare beneficiaries in Australia, with millions experiencing long delays waiting for their payments to be manually processed by Centrelink and Medicare. To address this, a recently elected government has made a campaign promise to introduce an **Algorithmic Decision-making System (ADS): a computer program that automates claim processing without human instruction**.

If effective, the ADS has the **potential to more efficiently process payments and reduce wait times for beneficiaries**. However, there are also concerns about the accuracy of the automated system and the possibility that, if not effective, it could lead to costly errors that would be avoided by human decision-makers.

Some researchers have expressed reservations about an immediate nationwide rollout, citing the potential for errors and lack of evidence proving the ADS is effective. They prepared a submission to the government proposing a randomised trial on a representative sample of Centrelink and Medicare claims from each state and territory in Australia. In this study, **half the claims would be processed by the new ADS (treatment group), while the other half would continue with the current method of manual processing by public servants (control group).**

The accuracy of the claims processed in both groups would be measured and compared. **The differences between the two groups would be used to determine the ADS's impact on the accuracy of claim processing**. The researchers contend that a wider rollout should only be pursued if the randomised trial demonstrates that claims are processed more accurately by the ADS than by public servants.

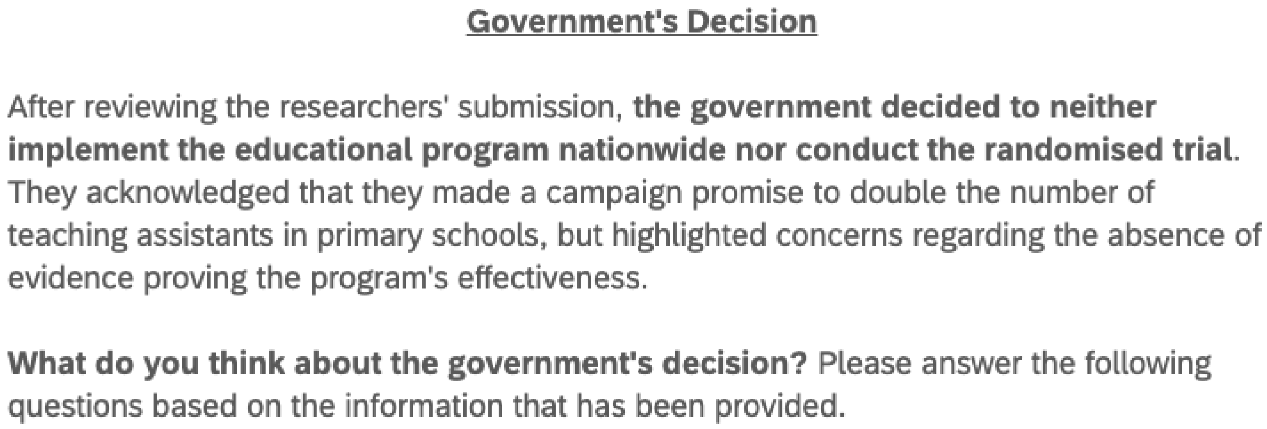
What do you think the government should do? Please rank the following options in order of your preference for what the government should do by moving your "most preferred" option to the top of the list ("1") and your "least preferred" option to the bottom of the list ("3").

Implement: Proceed with the nationwide rollout as initially planned 1

Experiment: Test the program with a randomised trial before deciding on the nationwide rollout 2

Neither: Do not proceed with the nationwide rollout and do not test the program with a randomised trial 3

Figure S11: Automated welfare claim processing

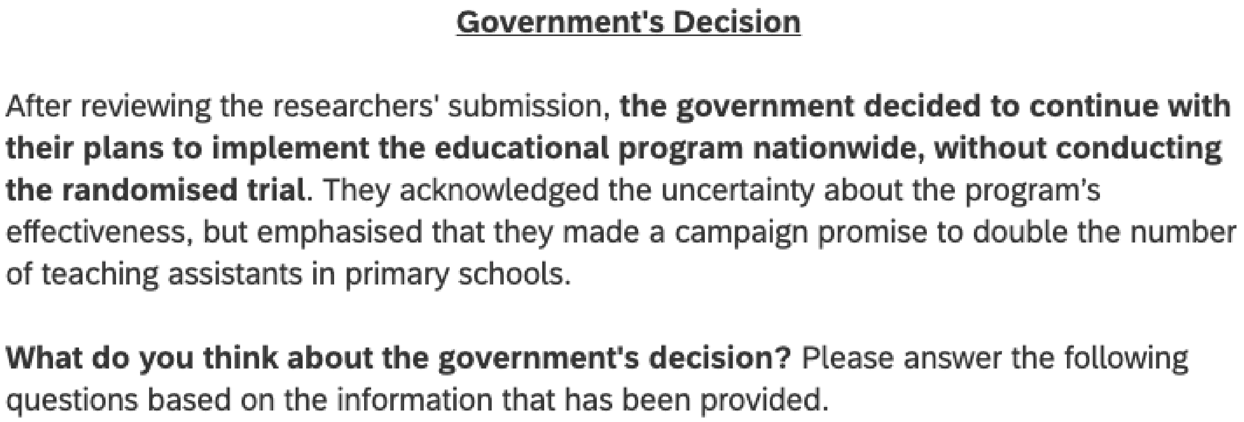


Government's Decision

After reviewing the researchers' submission, **the government decided to neither implement the educational program nationwide nor conduct the randomised trial**. They acknowledged that they made a campaign promise to double the number of teaching assistants in primary schools, but highlighted concerns regarding the absence of evidence proving the program's effectiveness.

What do you think about the government's decision? Please answer the following questions based on the information that has been provided.

Figure S12: Abandon in Experiment 1 (Education Policy)

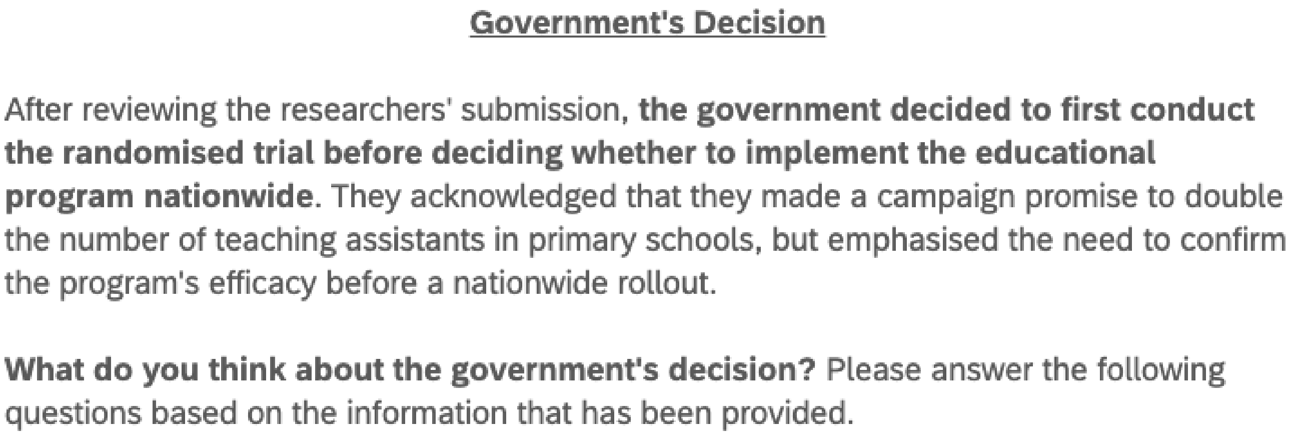


Government's Decision

After reviewing the researchers' submission, **the government decided to continue with their plans to implement the educational program nationwide, without conducting the randomised trial**. They acknowledged the uncertainty about the program's effectiveness, but emphasised that they made a campaign promise to double the number of teaching assistants in primary schools.

What do you think about the government's decision? Please answer the following questions based on the information that has been provided.

Figure S13: Implement in Experiment 1 (Education Policy)

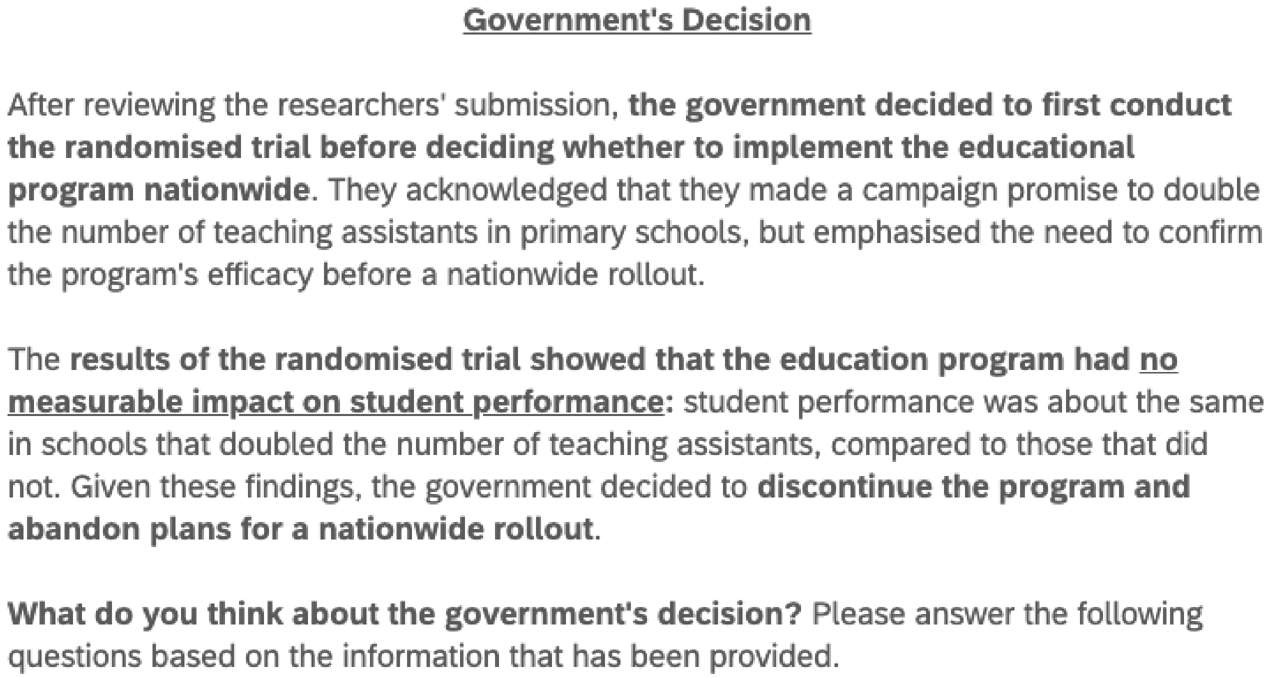


Government's Decision

After reviewing the researchers' submission, **the government decided to first conduct the randomised trial before deciding whether to implement the educational program nationwide**. They acknowledged that they made a campaign promise to double the number of teaching assistants in primary schools, but emphasised the need to confirm the program's efficacy before a nationwide rollout.

What do you think about the government's decision? Please answer the following questions based on the information that has been provided.

Figure S14: Experiment in Experiment 1 (Education Policy)



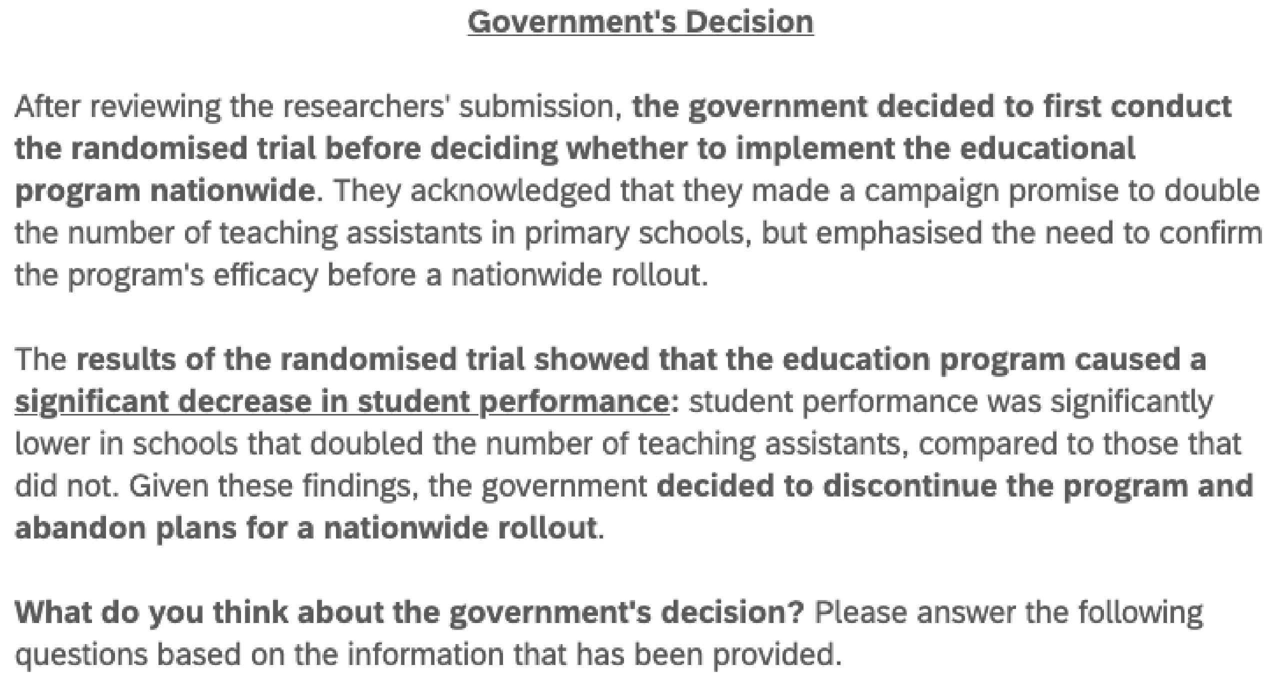
Government's Decision

After reviewing the researchers' submission, **the government decided to first conduct the randomised trial before deciding whether to implement the educational program nationwide**. They acknowledged that they made a campaign promise to double the number of teaching assistants in primary schools, but emphasised the need to confirm the program's efficacy before a nationwide rollout.

The **results of the randomised trial showed that the education program had no measurable impact on student performance:** student performance was about the same in schools that doubled the number of teaching assistants, compared to those that did not. Given these findings, the government decided to **discontinue the program and abandon plans for a nationwide rollout**.

What do you think about the government's decision? Please answer the following questions based on the information that has been provided.

Figure S15: No Impact in Experiment 2 (Education Policy)



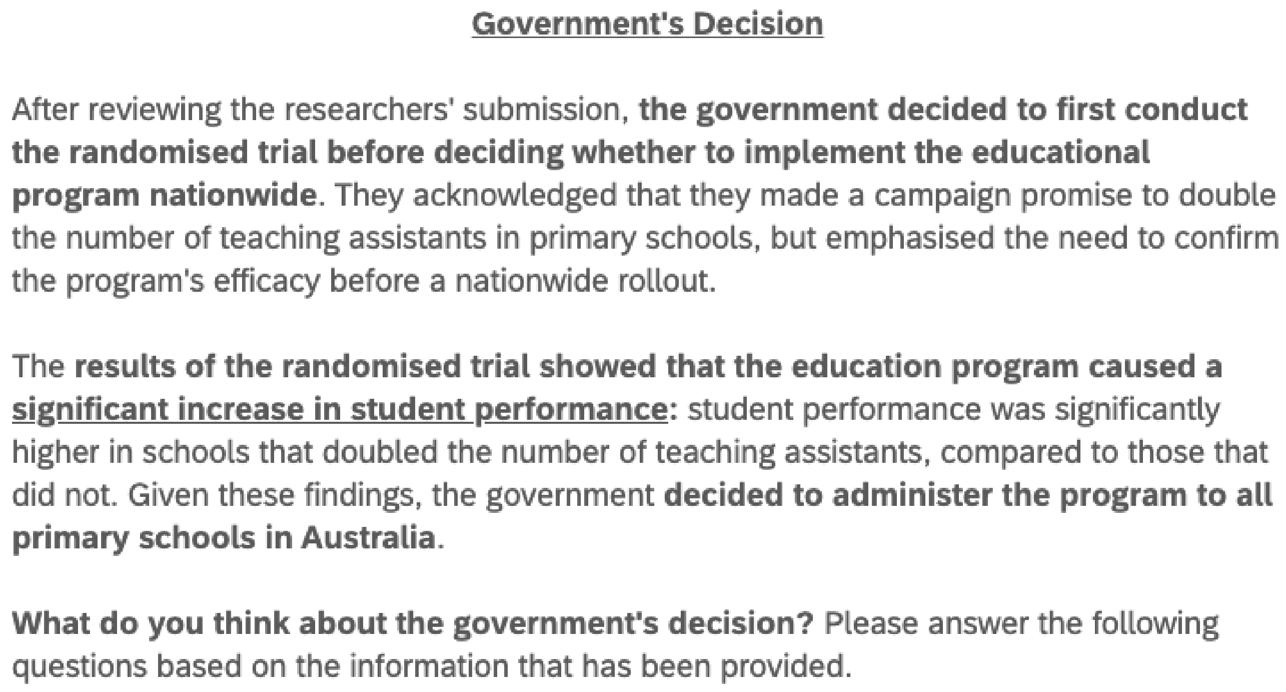
Government's Decision

After reviewing the researchers' submission, **the government decided to first conduct the randomised trial before deciding whether to implement the educational program nationwide**. They acknowledged that they made a campaign promise to double the number of teaching assistants in primary schools, but emphasised the need to confirm the program's efficacy before a nationwide rollout.

The **results of the randomised trial showed that the education program caused a significant decrease in student performance:** student performance was significantly lower in schools that doubled the number of teaching assistants, compared to those that did not. Given these findings, the government **decided to discontinue the program and abandon plans for a nationwide rollout**.

What do you think about the government's decision? Please answer the following questions based on the information that has been provided.

Figure S16: Positive Impact in Experiment 2 (Education Policy)



Government's Decision

After reviewing the researchers' submission, **the government decided to first conduct the randomised trial before deciding whether to implement the educational program nationwide**. They acknowledged that they made a campaign promise to double the number of teaching assistants in primary schools, but emphasised the need to confirm the program's efficacy before a nationwide rollout.

The **results of the randomised trial showed that the education program caused a significant increase in student performance:** student performance was significantly higher in schools that doubled the number of teaching assistants, compared to those that did not. Given these findings, the government **decided to administer the program to all primary schools in Australia**.

What do you think about the government's decision? Please answer the following questions based on the information that has been provided.

Figure S17: Negative Impact in Experiment 2 (Education Policy)

References

Anderson, Michael L. 2008. "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American statistical Association* 103(484):1481–1495.

Aronow, P.M., Jonathon Baron and Lauren Pinson. 2019. "A note on dropping experimental subjects who fail a manipulation check." *Political Analysis* 27(4):572–589.

Berinsky, Adam J, Michele F Margolis and Michael W Sances. 2014. "Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys." *American Journal of Political Science* 58(3):739–753.

Biddle, Nicholas, Matthew Gray and Michael Hiscox. 2023. "Public support for Randomised Controlled Trials and nudge interventions in Australian social policy.".

Gerber, Alan S., Donald P. Green and Edward H. Kaplan. 2014. The Illusion of Learning from Observational Research. In *Field Experiments and Their Critics*, ed. Dawn Langan Teele. New Haven: Yale University Press pp. 9–32.

Glass, Gene V. 1976. "Primary, secondary, and meta-analysis of research." *Educational researcher* 5(10):3–8.

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):18948–18950.

Kennedy, Brian and Meg Hefferon. 2019. "What Americans know about science.".
URL: *https://www.pewresearch.org/science/wp-content/uploads/sites/16/2019/03/PS_2019.03.28_science-knowledge_FINAL.pdf*

Lakens, Daniël. 2017. "Equivalence tests: A practical primer for t tests, correlations, and meta-analyses." *Social psychological and personality science* 8(4):355–362.

Levi, Margaret, Audrey Sacks and Tom Tyler. 2009. "Conceptualizing legitimacy, measuring legitimating beliefs." *American behavioral scientist* 53(3):354–375.

Lin, Winston. 2013. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique." *The Annals of Applied Statistics* 7(1):295–318.

Mazar, Nina, Christian T Elbaek and Panagiotis Mitkidis. 2023. "Experiment aversion does not appear to generalize." *Proceedings of the National Academy of Sciences* 120(16):e2217551120.

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):10723–10728.

Mislavsky, Robert, Berkeley Dietvorst and Uri Simonsohn. 2020. "Critical condition: People don't dislike a corporate experiment more than they dislike its worst condition." *Marketing Science* 39(6):1092–1104.

Montgomery, Jacob M, Brendan Nyhan and Michelle Torres. 2018. "How conditioning on posttreatment variables can ruin your experiment and what to do about it." *American Journal of Political Science* 62(3):760–775.

Peyton, Kyle. 2020. "Does trust in government increase support for redistribution? Evidence from randomized survey experiments." *American Political Science Review* 114(2):596–602.

Peyton, Kyle, Chagai M Weiss and Paige E Vaughn. 2022. "Beliefs about minority representation in policing and support for diversification." *Proceedings of the National Academy of Sciences* 119(52):e2213986119.

Peyton, Kyle, Gregory A. Huber and Alexander Coppock. 2021. "The Generalizability of Online Experiments Conducted During the COVID-19 Pandemic." *Journal of Experimental Political Science* pp. 1–16.

Peyton, Kyle, Michael Sierra-Arevalo and David G Rand. 2019. "A field experiment on community policing and police legitimacy." *Proceedings of the National Academy of Sciences* 116(40):19894–19898.

Rainey, Carlisle. 2014. "Arguing for a negligible effect." *American Journal of Political Science* 58(4):1083–1091.

Tanasoca, Ana and Andrew Leigh. 2024. "The democratic virtues of randomized trials." *Moral Philosophy and Politics* 11(1):113–140.

Ternovski, John and Lilla Orr. 2022. "A Note on Increases in Inattentive Online Survey-Takers Since 2020." *Journal of Quantitative Description: Digital Media* 2.

Vaughn, Paige E., Kyle Peyton and Gregory A. Huber. 2022. "Mass support for proposals to reshape policing depends on the implications for crime and safety." *Criminology & Public Policy* 21(1):125–146.