Journal of Experimental Political Science (2024), **11**, 343–359
doi:10.1017/XPS.2023.28

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Inaccurate forecasting of a randomized controlled trial

Mats Ahrenshop[1], Miriam Golden[2], Saad Gulzar[3] and Luke Sonnet[4]

[1]Department of Politics & International Relations, University of Oxford, Oxford, UK, [2]Department of Political and Social Sciences, European University Institute, Florence, Italy, [3]Department of Politics and School of Public and International Affairs, Princeton University, Princeton, NJ, USA and [4]Independent Researcher, Redwood City, CA, USA
Corresponding author: Saad Gulzar; Email: gulzar@princeton.edu

Abstract
We report the results of a forecasting experiment about a randomized controlled trial that was conducted in the field. The experiment asks Ph.D. students, faculty, and policy practitioners to forecast (1) compliance rates for the RCT and (2) treatment effects of the intervention. The forecasting experiment randomizes the order of questions about compliance and treatment effects and the provision of information that a pilot experiment had been conducted which produced null results. Forecasters were excessively optimistic about treatment effects and unresponsive to item order as well as to information about a pilot. Those who declare themselves expert in the area relevant to the intervention are particularly resistant to new information that the treatment is ineffective. We interpret our results as suggesting that we should exercise caution when undertaking expert forecasting, since experts may have unrealistic expectations and may be inflexible in altering these even when provided new information.

Keywords: Forecasting; RCTs; Experts; Information; ICT

Introduction

Randomized controlled trials (RCTs) constitute an invaluable method to estimate the effects of interventions on real-world behavior that is of interest to social scientists. The increasing use of RCTs intersects with larger transformations underway in the social sciences that involve greater commitments to reproducibility, transparency, and rigor. Taken together, RCTs, pre-registration, and pre-analysis plans have wrought a sea change in academic practices, including a drastic reduction in p-hacking and the development of more scientific research norms (Christensen and Miguel, 2018; Ofosu and Posner, 2023). Extending this commitment to increased rigor and transparency,

This article has earned badges for transparent research practices: Open Data and Open Materials. For details see the Data Availability Statement.

© The Author(s), 2023. Published by Cambridge University Press on behalf of American Political Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 06 Feb 2025 at 03:10:42, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/XPS.2023.28

scholars have recently begun to incorporate expert forecasting of experimental results into research practices to mitigate publication bias and improve experimental design as well as to clarify the interpretation of results (DellaVigna and Pope, 2019; DellaVigna, Otis, and Vivalt, 2020).

We study whether experts are appropriately skeptical of what RCTs can deliver. We investigate experimentally whether forecasters are sensitive to primes that we hypothesize ought to shift their expectations about the treatment effects of an intervention. Among other things, we specifically study whether subjects' forecasts change in light of new information that suggests the RCT may generate statistically insignificant results. We also consider heterogeneous effects among subjects who claim familiarity and who admit unfamiliarity with the area of research the RCT is concerned to assess.

The study recruited 280 experts from academic and policy research institutions in the United States and Pakistan, the country where the intervention we study was conducted. Our online forecasting experiment concerned the effectiveness of an RCT that deployed Information and Communications Technology (ICT) to improve political responsiveness (Golden, Gulzar, and Sonnet, 2023). In a 2×2 factorial design, we randomly vary the order of questions about compliance and intent-to-treat (ITT) effects, the provision of information that a pilot had taken place, and the provision of information regarding null pilot results. We thereby vary the amount of information subjects have available before making forecasts about compliance and treatment effects. We also ask subjects to update their ITT forecasts after revealing to them (null) pilot results.

We report three major results. First, subjects are generally unrealistically optimistic about treatment effects, and this is true even for subjects who are informed that a pilot study delivered null results. Second, randomizing the order in which subjects are asked to forecast *compliance* and *treatment effects* proves to have no significant effect on forecasts of treatment effects; that is, forecasters do not infer treatment effects based on their predictions about compliance rates, as we would expect. Similarly, varying whether subjects are exposed to the information that a *pilot* intervention generated null results has no significant effects on forecasts of treatment effects. In general, therefore, we do not find evidence that subjects incorporate information in meaningful ways into their forecasts and become more realistic about likely treatment effects. Our third main result is that only a subset of subjects – those who do not report expertise in the subject area of the RCT – update as expected on the basis of the pilot's null results. Thus, we provide evidence that researchers who self-identify as experts are especially unresponsive to information about the RCT.

The main take-away of the results we present is thus that academics and policy practitioners may too easily believe that an RCT is likely to be successful and may not be discouraged in these beliefs even when presented with information that should lead them to suspect an intervention will not generate effects. Moreover, subjects who consider themselves more expert in the subject matter domain of the intervention seem particularly resistant to incorporating new information into their forecasts. These results generate some concern that there might be too much systematic optimism about RCT interventions among forecasters. In our conclusions, we discuss how this could be taken into consideration when using forecasting as a tool to benchmark knowledge accumulation.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 06 Feb 2025 at 03:10:42, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/XPS.2023.28

Our findings contribute to the growing literature on forecasting. The appeal of forecasting to the social science community is evidenced by the creation of the Social Science Prediction Platform, a web-based forecasting platform that in its first ten months of existence attracted 1,700 users, 19 research projects seeking forecasts, and over 1,600 predictions (CEGA, 2021). Pioneering studies have used forecasts by experts to quantify the extent of novel information generated by particular RCTs (Casey et al., 2018; Cooper, 2018; Humphreys, Sánchez de la Sierra, and Van der Windt, 2019). Forecasts may prevent consumers of research results from discrediting evidence generated by an RCT with the claim that it corroborates preexisting expectations; the forecast clearly indicates what these expectations were in the first place. Thus, by forecasting an experiment prior to releasing its results, researchers put themselves on solid ground when they state that their intervention generated unexpected results. However, the work reported here suggests that forecasting does not represent a panacea. Instead, unrealistically optimistic forecasts may simply provide cover for experimental work that happens to succeed.

Experimental design and implementation

Sample

For our experiment, we recruited subjects from research institutions based in the United States and in Pakistan. Our subject pool was recruited via email as follows: 208 subjects from all individual members of the Comparative Politics Organized Section of the American Political Science Association; 35 subjects from all political science graduate students at two American universities (the University of California at Los Angeles and Stanford University); 27 subjects from policy professionals at a Pakistan research institute (the Centre for Economic Research in Pakistan), an organization whose staff has expertise in implementing randomized controlled trials. Finally, we also recruited 15 subjects in person in a graduate student seminar at an American university (the University of California at San Diego). Of the final sample of 280 subjects, 148 comprise university faculty, 77 graduate students, 16 post-doctoral fellows, and 39 other research staff. Forecasts were financially incentivized, with the top forecasting percentile receiving a modest payment depending on accuracy in relation to the actual RCT results. Information on the full composition of the experimental sample is reported in online Appendix, Table C2. Details regarding subject recruitment and Institutional Review Board approval are reported at the end of the article. The survey instrument is available as online Appendix E.

Summary of study for which forecasts were elicited

Subjects were asked to forecast the results of the experiment reported in Golden, Gulzar, and Sonnet (2023) that we briefly summarize now: the authors of the paper piloted and then scaled a field experiment where they partnered with Members of the Provincial Assembly (MPAs) of Khyber Pakhtunkhwa in Pakistan. They asked MPAs to use interactive voice response (IVR) technology to spur two-way communication between MPAs and heads of households who are their constituents. The treatment involved MPAs recording a message that was sent via robocall to

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 06 Feb 2025 at 03:10:42, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/XPS.2023.28

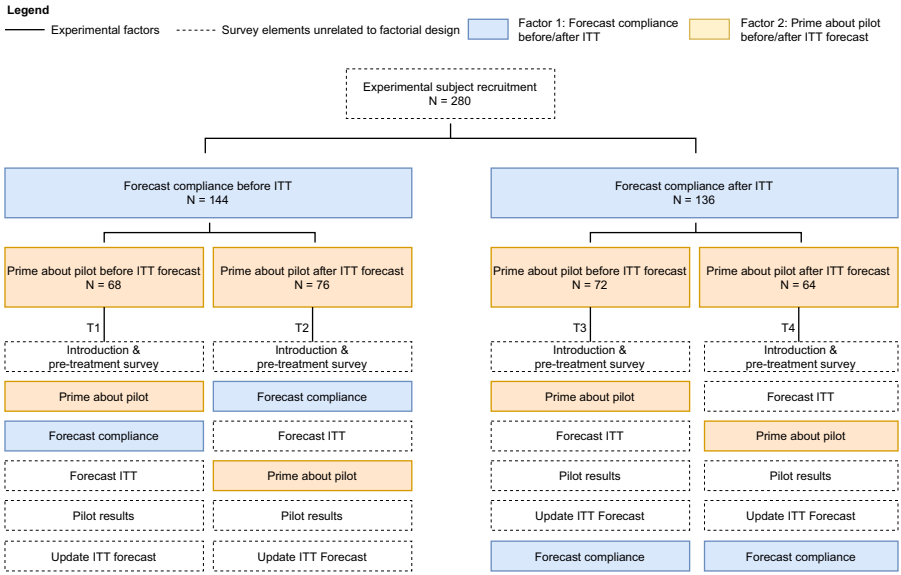

Figure 1. Experimental design. *Forecast compliance before/after ITT* refers to question ordering about when respondents forecast compliance rates and when they forecast treatment effects. Compliance percentages were requested for answering the MPA's call and for answering the MPA's question. Treatment effects were elicited in standard deviation units. *Pilot results* refer to the presentation of null treatment effects for three outcomes of interest in the pilot. *Prime about pilot* refers to a short informational vignette embedded in the survey where respondents read the following text: "Pilot and Scale-up: This program was designed through a pilot that we conducted with one MPA in [district name redacted] in 2016 (shown in red on the map). This pilot was conducted with 1,200 households. The scale-up project was implemented in the blue areas in this map."

constituents. Sometimes the message was accompanied by a question that constituents could answer by pressing keys on their phone.

In the analysis that follows, *Compliance* consists of the proportion of households who answered the phone, and when asked, the proportion who answered the question their MPA sent them. *Treatment effects* consist of whether treated households changed their *evaluations of their MPA*, their *evaluations of government performance*, and the *likelihood they would hold their MPA accountable for performance*. These outcomes are each constructed as an index drawing on multiple items (for details, see the notes in Table 3). In the text below, we refer to these as the MPA, GOVT, and ACCOUNTABILITY indices.

In addition to randomizing whether households received the initial *call* recorded by their MPA, the intervention randomized whether a follow-up phone call was *responsive* to the aggregate feedback received from voters or whether it was a *generic* follow-up. Results were null in both the pilot and the follow-up for all outcomes.

Hypotheses, allocation, and treatment

We assigned the 280 subjects to four different treatment arms formed by a 2 × 2 factorial design (see the experimental flowchart depicted in Figure 1) using simple

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 06 Feb 2025 at 03:10:42, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/XPS.2023.28

Table 1. Factorial experiment to elicit forecasts

		Forecast compliance	
		Before ITT	*After ITT*
Prime about pilot	*Before ITT*	T1 (N = 68)	T3 (N = 72)
	After ITT	T2 (N = 76)	T4 (N = 64)

random assignment. The forecasting survey was administered online via Qualtrics. Results of randomization tests, reported in online Appendix, Tables D1, D2, D3, and D4, present data confirming balance of treatment groups across measured covariates.

The forecasting survey employed a fixed set of primes and tasks for respondents. Experimental variation changed the order of the primes and tasks, thereby altering the amount and nature of information subjects had available at the time they made forecasts of treatment effects. The four arms generated by the 2 × 2 design differ only in the order with which certain primes and tasks about the RCT appear for survey participants. *Factor 1* consists of two conditions: whether subjects were asked to forecast the RCT compliance rate before or after they were asked to forecast the ITT effect. *Factor 2* consists of two conditions: whether subjects receive a prime about the fact that a pilot study of the RCT had taken place before or after the ITT forecast. We show this factorial design in a 2 × 2 matrix in Table 1.

This experimental randomization allows us to evaluate the following two hypotheses. First, we evaluate whether forecasting compliance before ITT effects *reduces* ITT estimates. The reasoning underlying this hypothesis is that field experiments in the domain of political participation are well known for low compliance. Field experiments about electoral turnout, for instance, generally elicit compliance rates of only about 25% (Schein et al., 2020), and those that use ICT to elicit political engagement generate compliance rates in the single digits (see Table A1 in Golden, Gulzar and Sonnet (2023)). Although high compliance rates by no means guarantee significant ITT effects, low compliance substantially undermines the likelihood that an intervention will generate significant treatment effects (see Gerber and Green (2012), ch. 5). Thus, asking subjects to consider compliance before asking them to forecast treatment effects in effect reminds them that compliance is likely to be problematic. We expect this item ordering to lower expectations for both compliance and treatment effects.

Second, we expect providing subjects the information that a pilot occurred will *increase* ITT estimates (and possibly compliance forecasts as well). Our reasoning is that funding agencies are more likely to allow scale-up if a pilot delivered significant results, whereas if a pilot "didn't work," scholars might abandon the project altogether (or redesign the study to improve outcomes). Thus, knowing that a pilot occurred should make subjects more optimistic about the outcome of the intervention.

The ordering of our primes and tasks, depicted in Figure 1, also permits experimental evaluation of a third hypothesis. Subjects who forecast compliance after ITT effects (T3 and T4) were also given the information that the pilot study did

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 06 Feb 2025 at 03:10:42, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/XPS.2023.28

not result in statistically significant findings.[1] We expect receiving information that pilot results were null will *decrease* forecasts of compliance. This is because the null pilot results could be interpreted post hoc as a result of poor compliance. However, it is possible instead that subjects expect the reverse: namely knowing a pilot produced null results and that a scale-up nonetheless occurred might lead subjects to imagine that the P.I.s undertook design modifications to prevent a recurrence of null results. If subjects interpret null pilot results in this fashion, then we expect that receiving information that pilot results were null will *increase* forecasts of compliance. Thus, the main hypotheses that we evaluate experimentally are as follows:

H1: Forecasting compliance before ITT effects results in lower ITT forecasts.

H2: Knowing a pilot occurred before forecasting ITT effects results in higher ITT forecasts.

H3a: Knowing a pilot produced null findings results in a lower compliance forecast.

H3b: Knowing a pilot produced null findings results in a higher compliance forecast.

We summarize the causal quantities of interest in Table 2.

Empirical strategy and estimation

We employ two strategies to construct estimates of the quantities of interest. Our main approach is to pool treatment arms according to Table 2, that is, depending on the outcome. For the ITT forecasts, we calculate difference-in-group-means via OLS of the form $Y_i = \alpha + \beta C_i + \varepsilon_i$ as well as $Y_i = \alpha + \beta P_i + \varepsilon_i$, with Y_i an individual's average forecast of the ITT (pooled across MPA, GOVT, and ACCOUNTABILITY indices). C_i (compliance prime) is 1 if an individual is in $\{T1, T2\}$ and 0 otherwise, and P_i (pilot prime) is 1 if an individual is in $\{T1, T3\}$ and 0 otherwise. This corresponds to estimating the quantities in rows 1 and 2 of Table 2, that is the effects of pilot and compliance primes on ITT forecasts.

For compliance forecasts, we calculate difference-in-group-means via OLS of the form $Y_i = \alpha + \beta R_i + \varepsilon_i$, where Y_i is an individual's compliance forecast estimated separately for two types of compliance (answering the MPA's call and then answering his question) and R_i is an indicator that takes a value of 1 if the individual was in $\{T3, T4\}$, that is, has seen the pilot results before making the compliance forecasts. This corresponds to estimating the quantity in the third row of Table 2, that is, the effects of pilot results on compliance forecasts.

An additional approach to the analysis of our ITT forecasts is to compare the different combinations of the 2×2 design by interacting the two treatment indicators representing the pilot prime first and the compliance forecast first, such that $Y_i = \alpha + \beta C_i + \gamma P_i + \delta C_i P_i + \varepsilon_i$. This allows us to compare those who (i) received the prime that a pilot had taken place as well as (ii) were asked to forecast compliance

[1]They were also given a number of other primes and tasks before forecasting compliance. We have no way to unbundle these from the information that pilot results were null, but theoretically we focus on the importance of the pilot results because we expect these to be especially consequential.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 06 Feb 2025 at 03:10:42, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/XPS.2023.28

Table 2. Estimands of interest

Hyp:	Comparison	Outcome	Effect of interest
H1	$\{T1, T3\}$ vs $\{T2, T4\}$	ITT forecast	Effect of pilot prime
H2	$\{T1, T2\}$ vs $\{T3, T4\}$	ITT forecast	Effect of compliance prime
H3	$\{T3, T4\}$ vs $\{T1, T2\}$	Compliance forecast	Effect of pilot results

Effect of pilot prime denotes the effect of receiving information that a pilot study took place prior to full RCT. *Effect of compliance prime* denotes the effect of being asked to forecast the compliance rate of the RCT prior to forecasting the ITT effect. *Effect of pilot results* denotes the effect of being asked to forecast the compliance rate after receiving the information that pilot results were null.

before making the ITT forecasts to those who (iii) were asked to forecast compliance after the ITT. This allows us to estimate whether the effect of the pilot prime on ITT forecasts depends on whether subjects were primed to think about compliance before making their ITT forecasts, a plausible conjecture as the hypothesized positive effect of a pilot prime on ITT forecasts could be offset by anticipation of low compliance. It also helps us understand the mechanism at work by teasing out whether the pilot prime also primes subjects to think about likely compliance, thereby illuminating the nature of the pilot prime treatment.

Summary statistics for all outcome measures and important pre-treatment covariates on which we condition in the analyses of CATEs are presented in online Appendix, Table C1.

Results

We begin with some descriptive statistics arising from the forecasts. In Table 3, we present realized treatment effects of the pilot study and of the scaled-up RCT along with average forecasts of these effects made by our subjects for different components of the experiment. In every case, forecast averages are higher than realized treatment effects. The only exception is the effect of the robocall (versus the control group that did not receive a call) on the ACCOUNTABILITY index, where the pilot produced treatment effects that were stronger than the intervention or that were forecast by experts. Moreover, mean forecasts are much larger than the ITT effects realized in the intervention, often more than five times as large. Thus, experts are unrealistically optimistic about the intervention, forecasting much stronger treatment effects than were in fact obtained.

The table also reports actual compliance rates of the intervention on the two major activities required: answering the MPA's phone call and, conditional on answering the phone, using IVR technology to answer a question that was posed by the MPA. The mean forecast for each was just over 50%, which substantially underestimates compliance in answering the phone but overestimates it for answering a question. This shows that the guesses made by forecasters about compliance rates are highly inaccurate and that experts are not able to predict compliance rates; their guesses are about 25 percentage points off the mark.

We now estimate two main outcomes of the forecasting experiment. The first is the forecast of the intervention's ITT. We elicit respondents' forecast of treatment effects in units of standard deviations. Forecasts were made for both the generic and

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 06 Feb 2025 at 03:10:42, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/XPS.2023.28

Table 3. Descriptive statistics of outcomes of interest in the pilot, RCT, and forecasts

Effect of ...	Pilot ITT	Intervention ITT	Mean forecast
Call on MPA	0.09 ($p = 0.518$)	0.02 ($p < 0.001$)	0.09
Call on GOVT	−0.05 ($p < 0.001$)	0.01 ($p < 0.001$)	0.06
Call on ACCOUNTABILITY	0.12 ($p < 0.001$)	0.00 ($p < 0.001$)	0.05
Responsive on MPA	0.01 ($p < 0.001$)	−0.02 ($p < 0.001$)	0.10
Responsive on GOVT	0.03 ($p < 0.001$)	0.02 ($p < 0.001$)	0.08
Responsive on ACCOUNTABILITY	0.05 ($p = 0.002$)	−0.01 ($p < 0.001$)	0.06
Compliance rate phone	NA	73.1 ($p < 0.001$)	48.9
Compliance rate question	NA	23.8 ($p < 0.001$)	50.9

Call refers to a comparison of the treatment arm in which households received a robocall containing a message and a question from their MPA versus a control group that received no call. *Responsive* refers to a comparison of the treatment arm in which households received a robocall containing a message and a question from their MPA and then received a responsive follow-up call (mentioning the action taken by the MPA in response to feedback) versus the arm which received a generic follow-up call (thanking citizens for input). MPA refers to an incumbent evaluation index that aggregates questions about (i) MPA feeling thermometer (1–10), (ii) MPA party feeling thermometer (1–10), (iii) voted for MPA (0/1). GOVT refers to a government evaluation index that aggregates questions about (i) state looks after me (1–5), (ii) importance of elections (1–5), provincial government competence (1–5). ACCOUNTABILITY refers to the prospects for accountability index that aggregates questions about (i) political efficacy (1–5), vote choice based on performance (1–6). *P*-values come from two-sided t-tests of the null hypothesis that the mean forecast is not equal to the pilot ITT or intervention ITT, respectively, hence indicating statistically significant differences between mean forecasts and pilot and intervention ITTs. Compliance rates are expressed as percentages. All other quantities report treatment effects in standard deviations.

responsive treatment arms of the RCT, and separately for effects of the treatment on indices of evaluations of the MPA, the government, and prospects for accountability in each arm (which we continue to label MPA, GOVT, and ACCOUNTABILITY). We also calculate an average across the three indices to retrieve an overall ITT forecast; this is the outcome reported in tables unless stated otherwise. We retrieved a 100% response rate with no attrition from the 280 subjects.

The second outcome of interest is respondents' forecasts of compliance rates in the RCT, both on (i) answering the MPA's call and then on (ii) answering a question conditional on (i). We solicit forecasts of compliance rates in percentages. There was minor attrition on these questions, with three out of the 280 subjects failing to answer.[2]

Forecasts of treatment effects

Results of the experiment on average ITT forecasts (pooling across MPA, GOVT, and ACCOUNTABILITY indices) are presented in Table 4. As is clear from the estimation results reported there, the two sets of forecasts we manipulated experimentally generate no significant treatment effects, either separately or

[2]As an additional robustness exercise, we present in Tables A.3, A.4 and A.5 treatment effects on forecasting errors (measured as absolute distance between the initial forecast and the actual compliance/pilot result) as well as the magnitude of updating (measured as the difference between revised and initial estimate) and direction of updating of forecasts (measured as the unit change in update toward correct estimate). We interpret this additional set of results as evidence that the informational treatments about the RCT design did not improve the accuracy of forecasts, nor did it lead respondents to update their forecasts, let alone in the correct direction.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 06 Feb 2025 at 03:10:42, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/XPS.2023.28

Table 4. Experimental effects for ITT forecasts (average across indices)

	Forecast call on ITT			Forecast responsive on ITT		
	(1)	(2)	(3)	(4)	(5)	(6)
Pilot prime	0.000		−0.005	0.004		0.002
	(0.006)		(0.009)	(0.007)		(0.009)
Compliance prime		0.009	0.003		0.010	0.008
		(0.006)	(0.009)		(0.007)	(0.009)
Interaction P × C			0.012			0.005
			(0.012)			(0.014)
Constant	0.065***	0.061***	0.064***	0.077***	0.074***	0.073***
	(0.004)	(0.004)	(0.007)	(0.005)	(0.005)	(0.007)
Num. obs.	280	280	280	280	280	280

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$. OLS estimates with HC2 standard errors. All forecasts are in standard deviation units of the treatment effects of the intervention. *Forecast call on ITT* refers to the comparison of the treatment arm in which households received a call with a message and a question from their MPA versus the control group that received no call. *Forecast responsive on ITT* refers to a comparison of the treatment arm in which households received a call with a message, a question, and a responsive follow-up call from their MPA (mentioning the action taken by the MPA in response to feedback) versus the arm that received a generic follow-up call (thanking citizens for the input).

interactively. Consider the results presented in row 1 of Table 4, which compares those who received the pilot prime before the ITT forecasts (T1 and T3) to those who received the pilot prime after making ITT forecasts (T2 and T4). The OLS estimates of this comparison appear in columns 1 and 4, the first for receiving the MPA's call and the second for later receiving responsive feedback. Knowing that a pilot occurred before being asked to make an ITT forecast does not alter forecasts in either direction; point estimates are essentially zero for all outcomes and treatment conditions.[3] The same is the case for the comparison presented in row 2, which pools ITT forecasts from T1 and T2 and compares them to the pooled forecasts from T3 and T4; this compares groups who made their compliance rate forecasts before forecasting ITT to those who forecast compliance rates after forecasting ITT effects. ITT forecasts are immune to change even when subjects are asked to think about compliance beforehand, as can be seen from the results reported in columns 2 and 5.[4]

Forecasts of compliance rate

We now turn to forecasts of compliance. Results are presented in Table 5, which reports pooled compliance forecasts from T3 and T4 and compares them to pooled

[3]Separate results for MPA, GOVT, and ACCOUNTABILITY are presented in Tables A.6 and A.7.

[4]The null results we present are substantively small as the forecasted treatment effect sizes are measured in standard deviation units. For example, the largest number in Table 4 corresponds to a 0.012 standard deviation increase in the forecast which is very small. Nevertheless, we also conduct a post-hoc power analysis reported in Table A.8 which shows that our study is well-powered to detect small effects in the range of approximately 0.02 standard deviation units, further allaying concerns that our inference is compromised by small samples.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 06 Feb 2025 at 03:10:42, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/XPS.2023.28

Table 5. Experimental effects for compliance forecasts

	Dependent variable:	
	Compliance rate for call	Compliance rate for question
Pilot results effect	−0.048**	−0.079***
	(0.024)	(0.029)
Constant	0.512***	0.547***
	(0.016)	(0.019)
Num. obs.	277	277

***$p < 0.01$; **$p < 0.05$; *$p < 0.1$. OLS estimates with HC2 standard errors. *Compliance rate for call* refers to the percentage of treated RCT subjects who answer the MPA's call. *Compliance rate for question* refers to the percentage of treated RCT subjects answering the call who also answer the MPA's question.

forecasts from T1 and T2. Subjects in T3 and T4 were presented a range of information about the pilot (including results), whereas subjects in T1 and T2 did not know pilot results before making compliance forecasts. The results are estimated via OLS for two forecast compliance rates (answering the call and then answering a question) on an indicator coded 1 if the subject is in $\{T3, T4\}$ and 0 if in $\{T1, T2\}$.[5]

Being provided a wide array of information about the pilot – in particular, knowing the pilot results were null – decreases forecasts of compliance rates by 5 percentage points for answering the MPA's call and 8 percentage points for then answering his question.[6] The same comparison shows that when subjects are asked to forecast compliance before forecasting treatment effects, the forecast compliance rate is significantly greater for both measures of compliance. When primed to think about compliance before giving an ITT forecast, the forecast compliance rates are 5 percentage points (for answering the call) and 8 percentage points (for answering a question) higher than the compliance rates predicted by subjects who were asked to forecast compliance at the end of the experimental module. Thus, forecasts of compliance appear sensitive to primes and to new information, unlike forecasts of treatment effects. We interpret this as evidence that negative shifts in perceptions of compliance induced by our information treatments do not automatically translate into lower – more pessimistic – estimates of the likely ITT. Thus, forecasters appear to disregard a crucial piece of information – compliance with the treatment in the RCT – when making forecasts about the likely effects of the intervention. This is particularly troubling because subjects' guesses about compliance are also highly inaccurate, as we have shown.

[5]As Figure 1 documents, another feature that differs between $\{T3, T4\}$ and $\{T1, T2\}$ is that subjects in the former treatment groups have also made their ITT forecasts before their compliance forecasts, which was not the case for the latter. However, we have no theoretically compelling reason to believe that being asked to forecast treatment effects would affect the forecast for compliance rates.

[6]Compliance forecasts were, however, inelastic to the pilot prime when we compare only T1 with T2. See Table A.1.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 06 Feb 2025 at 03:10:42, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/XPS.2023.28

Table 6. Effect heterogeneity for compliance forecasts

	Dependent variable:	
	Compliance rate for call	Compliance rate for question
Pilot results × Familiar	0.009	0.165***
	(0.048)	(0.058)
Pilot results	−0.056	−0.161***
	(0.034)	(0.040)
Familiar	0.048	−0.031
	(0.032)	(0.039)
Constant	0.490***	0.560***
	(0.023)	(0.026)
Num. obs.	276	276

***$p < 0.01$; **$p < 0.05$; *< 0.1. OLS estimates with HC2 standard errors. *Compliance rate for call* refers to the percentage of treated RCT subjects who answer the MPA's call. *Compliance rate for question* refers to the percentage of treated RCT subjects answering the call who also answer the MPA's question. *Familiar* refers to subjects who score themselves "familiar" or "very familiar" with ICT in governance.

Heterogeneity by expert status

Not all subjects are equally likely to use pilot results in their forecasts of compliance rates. In Table 6, we report heterogeneous effects for subjects who rated themselves on a familiarity scale ranging from 1–4, where 1 is least familiar and 4 is most familiar with the research domain. Subjects were asked, "How familiar are you with research on the use of information technology to improve governance?" The measure captures the subject's self-professed degree of expertise. We have dichotomized the familiarity variable so that subjects who responded (3, 4) on the familiarity question ("familiar" or "very familiar" with ICT in governance) are coded 1, and 0 otherwise. We report compliance rates for answering the call and then for answering the MPA's question conditional on having answered the call.[7]

Although the main effect of receiving information about pilot results on the compliance forecasts is statistically insignificant, the interaction effect for full compliance (answering a question conditional on having answered the call) is 16 percentage points higher for subjects who are more familiar with ICT research, as evidenced by the coefficient on the interaction term. That is, subjects who claim more expertise are more likely to forecast high compliance after seeing the null results of the pilot. (Note that we do not observe a similar conditional treatment effect for compliance rates with respect to answering the call.) This is exactly the reverse of subjects who admit unfamiliarity with ICT research. The latter forecast a similar 16 percentage point *decline* in full compliance once they see the pilot results. In other words, those familiar with the research area are not discouraged in their compliance forecasts by null results of the pilot; they remain equally optimistic

[7]We report (null) results for treatment effects interacted with respondents' optimism about information technology to improve governance in Table A.2.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 06 Feb 2025 at 03:10:42, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/XPS.2023.28

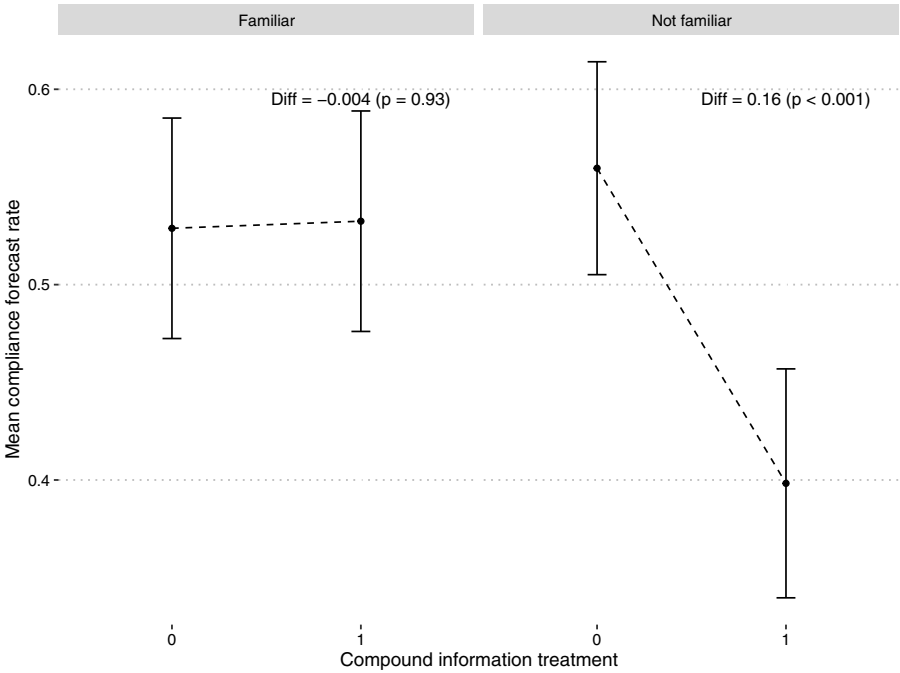

Figure 2. Heterogeneous treatment effects according to familiarity. We plot predicted compliance forecasts for different treatment groups according to whether they have received pilot results before making compliance forecasts or not separately for subgroups defined by familiarity with the use of technology in improving governance. To measure familiarity, we asked: "How familiar are you with research on the use of information technology to improve governance?" [1 = very familiar; 4 = very unfamiliar]. We dichotomized this variable by collapsing categories (1,2) into "familiar" and (3,4) into "unfamiliar."

about compliance rates even in the face of null results in the pilot. Thus, more expert forecasters are more inaccurate in their forecasts of compliance, conditional on seeing null pilot results.

Figure 2 breaks down this interaction effect into its constituent terms. The results depicted there show that subjects who rated themselves less familiar with information technology experiments were responsive to the treatment we induced; that is, when they saw the – largely discouraging – pilot results, they lowered their forecasts of compliance. Among those unfamiliar, the difference between the treatment and control group means of compliance forecasts is 16 percentage points and significant at $p < 0.001$; those in the treatment group have a 16 percentage point lower compliance forecast than those in control. By contrast, subjects who reported themselves familiar with information technology were less responsive to the information that the pilot generated null results and in fact did not alter their forecasts of compliance at all. Among this group of subjects, the difference between treatment and control is −0.4 percentage points and not significantly different from zero, again confirming that null results from the pilot did not depress their beliefs about compliance.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 06 Feb 2025 at 03:10:42, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/XPS.2023.28

Interpretation and discussion

We have reported three main results in the material above. First, expert forecasters expect the intervention to succeed, in the sense of generating significant treatment effects – although in fact, it did not. Second, item ordering for treatment effects and for compliance, inducing variation in the amount and nature of information available to forecasters, does not generate differences in treatment effect forecasts. Third, subjects familiar with the subject domain of the intervention are particularly unresponsive to new information and more overly optimistic than subjects who admit unfamiliarity with the subject domain. Expert subjects appear wedded to the belief that an intervention will produce significant treatment effects despite evidence that this may not be the case.

Is this apparent optimism about RCTs warranted? The data suggest not. In economics, about half of RCTs produce null results, according to multiple analyses of published results (see Snyder and Zhuo (2018), tab. 2 and Brodeur, Cook, and Heyes (2020), Fig. 2). In impact evaluations of development programs, most published results are likewise statistically insignificant (Vivalt, 2019). In political science, we collected data on all RCTs published in the discipline's top three journals between 2012 and 2021 (through the April 2021 issues). We report the data in Figure 3, which depicts bar graphs showing the proportion of articles in the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics* that feature RCTs, and of these, the proportion with null results on the main treatment effect. The data show that no more than 5% of articles in the top political science journals feature RCTs over the past ten years, revealing the novelty of RCTs in the political science discipline. However, the proportion of published null results within those RCT studies is high: nearly 50% for the *APSR* and around 33% for *AJPS* and *JOP*. Thus, readers of RCTs in economics and political science are accustomed to seeing null results in the most prestigious and visible disciplinary journals.

Of course, large numbers of null results in RCTs might be what we should expect. It is difficult to p-hack results of a randomized controlled trial; indeed, it is in part for this very reason that RCTs constitute the scientific benchmark. But RCTs are not themselves drawn from a random distribution. That is, scholars do not perform field experiments randomly selecting from all possible experiments and locations. Instead, it is a reasonable guess that scholars select experiments they *hope* will be successful. Thus, we might anticipate more than 5% of RCTs to produce significant results at the 5% level. But how much more? We have no way to know. Despite this inbuilt bias favoring significant results, the average published RCT is statistically likely to generate a null result, as the data depicted in Figure 3 document. Thus, the academics and policy practitioners who participated in our forecasting experiment are on average unrealistically optimistic about the treatment effects that an RCT is likely to generate. If a respondent has no specific expertise or insight suggesting otherwise, the "right" answer about whether the average RCT will succeed is that it will generate a treatment effect of 0. The overoptimism that we observe is unlikely to be the result of publication bias because consumers of RCTs have been exposed to many published null results.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 06 Feb 2025 at 03:10:42, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/XPS.2023.28

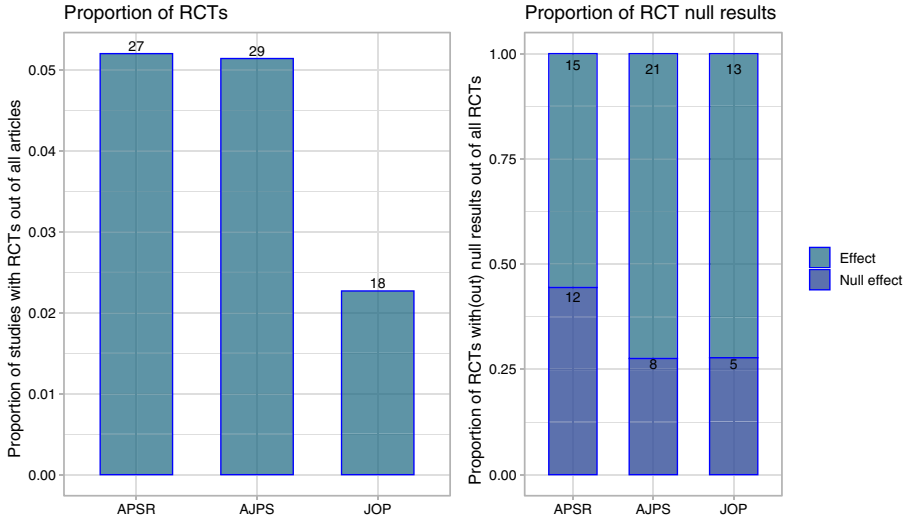

Figure 3. Published RCTs in political science and those reporting null results, 2012–2021. The left-hand panel shows the proportion of articles reporting results from an RCT out of all articles published in the respective journal. Over the period, the *APSR* published a total of 519 articles, *AJPS* 564, and *JOP* 793. The number of articles with RCTs is shown at the top of each bar. The right-hand panel looks only at articles that reported results from an RCT and distinguishes between those that reported null results on the main treatment effect of the intervention and those that reported a significant treatment effect. Relevant numbers are shown in each bar portion. Data collection procedures and coding principles detailed in online Appendix B.

Possibly, the experts we enrolled in the forecasting experiment read the description of the RCT that was presented and thoughtfully concluded that it was specifically likely to generate significant results. But it is difficult to interpret our data that way. Respondents more familiar with the subject domain of the intervention were more resistant to using new information to recalibrate their forecasts than inexpert respondents. This suggests that expertise may carry with it particular bias and that experts in information treatments to improve governance are deeply wedded to the hope that these treatments will be effective.

Could subjects who profess expertise be more likely to know the IVR P.I.s either personally or by reputation and to believe that their work is particularly likely to produce significant results? In the data we have, there is no evidence that reputational considerations were in play. Examining the composition of subjects who profess expertise finds that they are relatively evenly distributed across professional categories (faculty, post-doctoral fellows, graduate student, research staff) and across experimental sites (the Pakistani research institute, the APSA CP list of members, and graduate seminars), as we show in Tables 7 and 8. If reputational effects were in play, arguably research staff in Pakistan and faculty members would be more familiar with the reputations of the P.I.s and thus more likely to defer to their expertise. But this is not the case.

The real test will come when more data about social science predictions are available from the Social Science Prediction Platform, and it becomes possible to

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 06 Feb 2025 at 03:10:42, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/XPS.2023.28

Table 7. Distribution of subjects professing expertise by professional status

Expertise	Faculty	Grad student	Post-doc	Research staff	Total
Unfamiliar	83	40	9	9	141
Familiar	65	36	7	30	138
Total	148	76	16	39	279

Table 8. Distribution of subjects professing expertise by site

Expertise	Research institute	APSA CP members	Grad students	Classroom	Total
Unfamiliar	7	108	19	7	141
Familiar	20	95	15	8	138
Total	27	203	34	15	279

assess whether forecasting is generally over-optimistic for RCTs conducted by more prominent scholars.

In the meantime, our findings lend additional weight to other work, in particular (Dunning et al., 2019, ch. 12), that shows that policy practitioners continue to support funding research even when a meta-analysis reports that the research agenda is unlikely to be successful. That is, experts seem reluctant to update even in the face of evidence that suggests greater caution is warranted for a particular line of work. We applaud caution in interpreting single studies and agree that multiple pieces of research should underlie shifts in expectations. But we would like to see researchers adjust their expectations in light of cumulative findings that point one way or another, and in the meantime, we contend that researchers should be more circumspect in their forecasts.

Statements

Ethics statement

The research reported in this paper was approved under expedited review procedures by the Institutional Review Board of the University of California at Los Angeles, IRB#17-000182-AM-00005 on 31 January 2019.

The subject pools were recruited via email from all individual members of the Comparative Politics Organized Section of the American Political Science Association, from all political science graduate students at the University of California at Los Angeles (UCLA) and Stanford University, and from all policy professionals at the Centre for Economic Research in Pakistan. (CERP), an organization whose staff has expertise in implementing randomized controlled trials. All subjects were provided informed consent and were free not to take the survey or to withdraw at any time. Members of the Comparative Politics Organized Section of the American Political Science Association were offered Amazon gift

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 06 Feb 2025 at 03:10:42, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/XPS.2023.28

cards in the amount of 10 USD if their answers put them in the top quarter of respondents in that subject pool. Graduate students at UCLA and Stanford University were offered Amazon gift cards in the amount of 10 USD if their answers put them in the top quarter of respondents in each subject pool. CERP professionals were offered phone credits in the amount of 1,000 PKR if their forecasts were within 0.2 standard deviation units (on average) of the actual IVR results.

Additional subjects were recruited at a seminar at New York University and among undergraduates at the University of Peshawar. Forecasts by these subjects are not included in the analysis, because at these sites, the survey was administered on paper and no randomization took place.

Forecasts were collected on Qualtrics.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/XPS.2023.28

Data availability. The data, code, and additional materials required to replicate all analyses in this article are available at the *Journal of Experimental Political Science* Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/W2ITZG.

Funding. Funding for the research reported here was provided by Stanford University to Saad Gulzar and by the Governance Initiative of the Abdul Latif Jameel Poverty Action Lab (J-PAL), #300049 to Miriam A. Golden, Saad Gulzar, and Luke Sonnet. Mats Ahrenshop gratefully acknowledges PhD funding from the ESRC and Trinity College Oxford.

Competing interests. There are no conflicts of interest.

References

Ahrenshop, Mats, Miriam A. Golden, Saad Gulzar, and Luke Sonnet. 2023. "Replication Data for: Inaccurate Forecasting of a Randomized Controlled Trial." *Harvard Dataverse*. https://doi.org/10.7910/DVN/W2ITZG

Brodeur, Abel, Nikolai Cook, and Anthony Heyes. 2020. "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics." *American Economic Review 110*(11): 3634–60.

Casey, Katherine, Rachel Glennerster, Edward Miguel, and Marten Voors. 2018. *Skill Versus Voice in Local Development*. Working Paper 25022. Cambridge: National Bureau of Economic Research (NBER).

CEGA, (Center for Effective Global Action). 2021. "Emerging Benefits and Insights from a Year of Forecasting on the Social Science Prediction Platform." https://medium.com/center-for-effective-global-action/emerging-benefits-and-insights-from-a-year-of-forecasting-on-the-social-science-prediction-platform-1f554e850a57

Christensen, Garret, and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature 56*(3):920–80.

Cooper, Jasper. 2018. "How Robust is Institutionalized Corruption? A Field Experiment on Extortion along West African Highways." Unpublished paper.

DellaVigna, Stefano, and Devin Pope. 2019. *Stability of Experimental Results: Forecasts and Evidence*. Working Paper 25858. Cambridge: National Bureau of Economic Research (NBER).

DellaVigna, Stefano, Nicholas Otis, and Eva Vivalt. 2020. "Forecasting the Results of Experiments: Piloting an Elicitation Strategy." *AEA Papers and Proceedings 110*:75–79.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, Craig McIntosh, and Gareth Nellis, eds. 2019. *Information and Accountability, and Cumulative Learning: Lessons from Metaketa I*. Cambridge: Cambridge University Press.

Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton.

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 06 Feb 2025 at 03:10:42, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/XPS.2023.28

Golden, Miriam, Saad Gulzar, and Luke Sonnet. 2023. "'Press 1 for Roads': Descriptive and Experimental Evidence on Political Communication." Unpublished paper.

Humphreys, Macartan, Raúl Sánchez de la Sierra, and Peter Van der Windt. 2019. "Exporting Democratic Practices: Evidence from a Village Governance Intervention in Eastern Congo." *Journal of Development Economics 140*: 279–301.

Ofosu, George K. and Daniel N. Posner. 2023. "Pre-Analysis Plans: An Early Stocktaking." *Perspectives on Politics 21*(1): 174–90.

Schein, Aaron, Keyon Vafa, Dhanya Sridhar, Victor Veitch, Jeffrey Quinn, James Moffet, David M. Blei and Donald P. Green. 2020. "A Digital Field Experiment Reveals Large Effects of Friend-to-Friend Texting on Voter Turnout." SSRN Working Paper. https://ssrn.com/abstract = 3696179, https://doi.org/10.2139/ssrn.3696179.

Snyder, Christopher and Ran Zhuo. 2018. Sniff Tests in Economics: Aggregate Distribution of Their Probability Values and Implications for Publication Bias. Working Paper 25058. Cambridge: National Bureau of Economic Research (NBER).

Vivalt, Eva. 2019. "Specification Searching and Significance Inflation across Time, Methods and Disciplines." *Oxford Bulletin of Economics and Statistics 81*(4): 797–816.

Cite this article: Ahrenshop M, Golden M, Gulzar S, and Sonnet L (2024). Inaccurate forecasting of a randomized controlled trial. *Journal of Experimental Political Science* **11**, 343–359. https://doi.org/10.1017/XPS.2023.28

Downloaded from https://www.cambridge.org/core. Berklee College Of Music, on 06 Feb 2025 at 03:10:42, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/XPS.2023.28