Cumulative knowledge in the social sciences: The case of improving voters' information*

Federica Izzo Torun Dewan Stephane Wolton

Current draft: May 24, 2020

Link to the most recent version

Abstract

Cumulative knowledge requires (at least) two conditions to be met: unbiasedness and comparability. Research designs should be unbiased so that researchers obtain correct estimates of an underlying quantity. Empirical specifications, the actual regression run, should permit comparability so that researchers measure the same quantity across distinct studies. The first condition is covered by the causal revolution, the second is the object of this paper. Using the example of interventions providing additional information to voters, we show that unbiasedness does not imply comparability. Any two studies that employ the commonly used specification to analyze the electoral consequences of informational campaigns estimates different estimands. This holds true even after removing all external validity issue, all statistical noise, and all sources of bias. We highlight conditions to restore comparability and describe specifications that satisfy them.

Key words: electoral accountability, reproduction, accumulation of knowledge, comparability, bias, good news, bad news

*We thank Matilde Bombardini, Peter Buisseret, Selina Hofstetter, Ehtan Kaplan, Mark Kayser, Mark Meredith, Rocio Titiunik, Janne Tukiainen, and conference participants at the 2018 Columbia PE Conference and 2019 IAST Political Economy Conference for their helpful comments and advice. All remaining errors are the authors' responsibility. Corresponding author: Federica Izzo fizzo@ucsd.edu

1 Introduction

How do we know in the social sciences? We see two broad perspective on this question, which, at the risk of over-simplifying, can be summarized as such. The first view emphasizes the role of theory. We know when we have a convincing explanation for a well-documented pattern. The second stance does not deny the importance of theory, but takes a more pragmatic approach, relying more on measurements and empirical observations. Epitomized by the JPAL lab in economics or the path-breaking Metaketa Initiative in political science, this second persepctive seeks to "generalize upon sound experiments, draw analogies, and build up scientific conclusions" (Hacking, 1983: 114). Or, to quote recent Nobel prize winners Abhijit Banerjee and Esther Duflo (2009, page 161), it builds on the idea that "cumulative knowledge is generated from related experiments in different contexts."

We are sympathetic with this second approach that stresses intervening (experiments) on top of or beyond representing (theory), to paraphrase Hacking (1983). In this paper, however, we contest the implicit assumption that the accumulation of knowledge only requires unbiased research designs. Absence of bias only guarantees that the results of a specific study are a correct estimate of a certain quantity. While necessary, unbiasedness is not sufficient to uncover stable phenomena. For a sequence of measurements to converge to the same value, all analyses should estimate the same parameter, the same estimand. In other words, studies should be comparable. Absent comparability, researchers may attribute varying estimates across different studies to a lack of external validity, abandoning hopes to establish a stable phenomena, when it may well be due to all studies measuring different quantities. And this may spillover into the wrong advices given to policy-makers. Comparability is, thus, a critical component of knowledge accumulation. And to check whether it is achieved, you need theory we argue.

To understand our notion of comparability and how it differs from external validity or heterogeneous effect, it helps first to introduce a distinction between the context and the circumstances in which a study takes place. The context corresponds to the fundamental attributes of (say) a country. It captures the distribution of politicians' characteristics, the average education of voters, the usual resources available to office-holders (e.g., to carry out public good projects), the baseline economic conditions in the country. In social sciences, researchers draw a set of observations at a particular point in time, possibly in a certain region or district within the country. These observation are characterized by a given realization of these attributes. We, thus, denote the circumstances the *specific* conditions in which the intervention takes place (e.g., the state of the economy at the time of the study).

The quantity estimated in any empirical study is always a function of the context, and, sometimes, of the circumstances as well. In this latter case, when analyses yield estimates of circumstances-specific estimands, the search for stable phenomena is severely impaired. By definition, the circumstances are time-and-place specific: no two studies will ever be conducted under the same circumstances and, consequently, no two studies will ever measure the same quantity. Knowledge accumulation becomes hopeless. When researchers run specifications that purge circumstances from the quantities they estimate, estimands become solely a function of the context. As a result, fixing the fundamentals, any two surveys employing the same regression measure the same quantity, even if they take place in different circumstances. This is precisely our notion of comparability. Discussion of external validity—how well findings travel from one context to the next—only makes sense when studies are comparable.

We illustrate our contention with the example of informational campaigns and their consequences for electoral accountability. There are several reasons for this choice. First, there exists a large body of theoretical work on the effect of information on voters' (and politicians') behaviors (see, among many others, Ashworth, 2012; Ashworth et al., 2017). Second, the question has attracted considerable empirical interest. It is, for example, the theme of the first set of coordinated studies under the Metaketa Initiative (Dunning et al., 2020). Third, the topic is especially interesting because empirical research, despite its many measurements, has not reached a definitive conclusion on the electoral effect of providing information to voters. As Bhandari, Larreguy, and Marshall (2019: 2) explain, "recent studies identifying the effects of informational campaigns on electoral accountability (...) yield mixed findings."

There exist many explanations for this lack of consistent results. Two are commonly advanced: statistical noise and lack of external validity. The first issue can be mitigated by multiplying studies. The second is more problematic for cumulative knowledge. It suggests that the impact of informational campaigns varies significantly from one context to the next so that, as we noted, there is no hope of uncovering a stable phenomenon. We suggest that we should not reach this conclusion too fast. Indeed, our paper proposes a third rationale for these mixed findings: lack of comparability across studies.

In our analysis, we suppose that researchers have access to an infinite number of observations, rendering the issue of statistical noise moot. We eliminate sources of bias by assuming that researchers randomly sample a mass of observations before randomly dividing them between treated and control units. We also assume that external validity is not a concern, since all studies draw samples from the same setting, with the same underlying fundamentals. We finally impose that all studies employ the same empirical specification. Consistent with many works in the literature, researchers condition the effect of informational campaigns on good news (information that is meant to raise the voters' evaluation of office-holders) and bad news (information that is supposed to decrease the voters' opinion of their incumbent). That is, they compare outcomes in treated units with outcomes in the control units which would have received the same information if treated. We, however, maintain that each sample draws distinct observations. Researchers observe the consequences of their informational campaign under different circumstances, such as specific time-varying economic conditions. And this is enough to render two studies with a well-planned research design non-comparable.

To see why, consider the following thought experiment, which matches our theoretical framework below. Two studies analyze the consequences of informational campaign revealing instances of corruption, or lack thereof, by office-holders to voters. These two studies are conducted in the same country (context) at two different points in time. And in this country, voters put significant weight on politician's honesty in their voting decision as honest politicians are more likely, everything else equals, to provide public goods to their constituents. The key difference between the two studies is the economic circumstances at the time of the intervention. The first study occurs in favorable times (e.g., harvest has been good or prices of natural resources the country export is high). The other, instead, happens in time of relative hardship (perhaps due to unexpected weather events or negative shocks to price of raw materials).

How does it affect the evaluation of the informational campaign? In the first case, money flows to office-holders allowing them to realize several development projects, even when they happen to be corrupt. In the second case, the budget is much smaller and the number of projects completed decreases substantively, even for honest incumbents. Absent perfect information about the economic environment, this means that, in the first survey, the office-holder's performance is high and many voters have a relative good opinion of their representative. In the second study, incumbents' performance is low and many constituents evaluate poorly their representative. Think now about the effect of providing good news, say pieces of information that indicate or reveal that the incumbent is honest. In the first study, this moves many voters' posteriors slightly upward from a relatively high baseline. As a large proportion of treated constituents barely changes their evaluation, comparing treated and control units, researchers uncover a low effect of the informational campaign. In the second study, good news moves many voters' evaluation upward from a low baseline. As the opinion of a large proportion of treated constituents significantly improves, researchers recover a large effect of the informational campaign. Everything looked the same: randomization was perfect in the two studies, researchers analyze the same intervention, they run an identical regression. And yet, they uncover estimates of different, circumstance-specific estimands.

The scope of the comparability issue we uncover is potentially quite large. To avoid it with certainty, it is necessary that researchers condition on *all* electorally relevant information voters obtain absent intervention. This involves a subtle change of counterfactuals. Instead of asking what if the control group were to be treated (as researchers implicitly do by conditioning on good or bad news), researchers should ask what if the treated units had not received treatment. Instead of conditioning on what control units would have observed if treated, researchers should condition on what the treated villages would have learned absent treatment. In practice, this corresponds to running a fully saturated model with interactions between the treatment (good or bad news) and what constituents learn about their incumbent absent researchers' interventions. Doing so, researchers can recover specific theoretical quantities of interest (note the plural): the pure effect of information relative to a level of performance. Each estimand is solely a function of the context, none of them is affected by the precise circumstances in which the study takes place, and comparability is obtained.

This recommendation may be difficult to put into practice in certain circumstances. Indeed, it requires precise knowledge of the local circumstances if the unit of analysis is villages or municipalities (as in Ferraz and Finan, 2004). And any omitted piece of information may be enough to break down comparability across studies. Instead, we show how comparability is easily restored if the unit of analysis is individual voters. Then, researchers just need to collect constituents' opinion of the incumbent prior to the (possible) provision of information and condition their analysis on voters' evaluation. Indeed, a well known property of constituents' belief about their representative is that it encompasses all the relevant information available to voters. That is, they deal with the exact problem our work identifies.

While our paper is mostly about the scientific process, it is not without policy relevance. Think about a group of advisers who canvass the literature before deciding whether to advise politicians about implementing informational campaigns. If these aides are interested in improving electoral accountability, the mixed findings existing papers find would probably bias them against recommending such interventions. We, in turn, show how our approach can give upper and lower bounds on the electoral effects of providing information to voters. And these limits are comparable across studies within a given context, and so, possibly, across countries.

Overall, we believe that the empirical literature in social sciences has made considerable progress towards the accumulation of knowledge. The focus on unbiased studies triggered by the causal revolution is an absolutely necessary condition to learn about social phenomena. But it is not a sufficient one. Cumulative knowledge also requires comparability across studies. And this paper, we hope, advances our understanding of when empirical works measure the same quantity, and when they do not. In what follows, we present our arguments in greater details in two forms. In the main text, we use a formal model. In Appendix A, we develop our reasoning with the help of the potential outcome framework. Each approach can be read separately, but both are complementary in our view.

2 Literature review

At a conceptual level, our work contributes to the recent but burgeoning literature that uses formal models to connect theoretical and empirical counterfactuals, an approach referred to as theoretical implications of empirical models. For example, Eggers (2017) unpacks the properties of regression discontinuity designs, Ashworth et al.'s (2018) parses out possible mechanisms for why random events may affect incumbents' electoral fortune, Bueno de Mesquita and Tyson (2019) highlights the difficulty to recover theoretical quantities of interest when the dependent variables and the main explanatory factors are the results of strategic behaviors (e.g., the effect of protests, whose size depends on citizens' strategic decisions, on repression, a policy-maker's strategic choice). The closest paper to ours in this burgeoning literature is Wolton (2019) who details how estimates of the electoral consequences of biased media are a function of the political environment (e.g., partisan identity of office-holder) and the media environment (e.g., the partisan identity of media reporting). Hence, Wolton (2019) is interested in the external validity (or lack thereof) of estimates across different contexts. In contrast, we look for conditions under which empirical studies recover the same estimand fixing the context in which researchers intervene. We also propose a theoretical framework which includes many voters, villages, and districts. This allows us to precisely map equilibrium outcomes into empirical quantities of interests.

From a more substantive standpoint, our paper relates to the large body of theoretical work on the effect of information. Following a long tradition (see Ashworth, 2012), we use a political agency framework to model the relationship between voters and their representatives. More precisely, our work connects with a host of papers studying the impact of providing additional information to voters (e.g., Prat, 2005; Fox, 2007; Ashworth and Shotts, 2010; Fox and Van Weelden, 2012; Ashworth and Bueno de Mesquita, 2014). However, while previous studies focus on the normative consequences of greater transparency, we are concerned with the empirical analysis of informational campaigns.

A recent work by Grossman, Michelitch, and Prato (2018) studies a political agency model with features analogous to ours (for example, both assume that politicians' performance is binary and, thus, a coarse signal of their underlying ability or honesty). The objectives, though, are fundamentally different. Grossman, Michelitch, and Prato (2018) focus mostly on the empirical implications of their theoretical model. They use their formal framework to generate novel predictions on the effect of information campaigns, and then test these predictions against empirical data. In turn, the objective of our paper is to use a formal model to establish theoretical properties of existing empirical estimates, and identify conditions under which comparing the results of different studies permits knowledge accumulation. To the best of our knowledge, our paper is unique in this respect in the formal literature.

3 Set-up

In this section, we first describe our theoretical set-up, which consists of a political agency model with adverse selection slightly twisted to produce theoretical equivalents of empirical estimates of interest. We then explain how we model interventions. After detailing the timing, we conclude this section with a few remarks on our modeling choices.

Basics

We consider a country divided into D districts with one representative in each. Each district is constituted of V villages, each inhabited by a mass 1 of voters. With some small abuse of notation, we use D, and V to denote both the set of units and the cardinality of this set.

The model has two periods. At the beginning of the game, an incumbent (I_d) is in office in each district $d \in D$. After the first term in office, each incumbent is up for re-election: all voters who live in district d cast a ballot for either the incumbent I_d or a randomly drawn challenger C_d.

In each of the two periods, a public project needs to be completed in each village v. For example, a road must be repaired, a school must be built, or a hospital must be completed. Each village-specific project is either completed/successful ($\omega_v^t = 1$), or non-completed/unsuccessful ($\omega_v^t = 0$). The probability that a project is completed depends solely on (i) the the office holder's type and (ii) the underlying economic circumstances.

Each politician is one of two types: honest or corrupt ($\tau \in \{h, c\}$). Villagers do not know their incumbent's type at the beginning of the game. However, it is common knowledge that the incumbent is honest with probability π^I and the challenger with probability π^C (identical in all districts for simplicity). We assume that honest types are always more likely to succeed in completing a project than corrupt ones, for example due to embezzlement and diversion of funds by the latter.

Economic factors, which, for example, influence the size of the budget available to office-holders, are modeled as a shock (formally, state of the world), denoted by θ^t. We assume that θ^t is common to all office-holders (without much loss of generality). Villagers do not observe the realization of θ^t, but it is common knowledge that this shock is independently and identically distributed each period according to the continuous and strictly increasing cumulative distribution function F over the interval $[\underline{\theta}, \overline{\theta}]$, with $\underline{\theta} < \overline{\theta}$.

We assume that the probability a village-specific project is successful takes the form $Pr(\omega = 1) = \alpha(\tau) \times \gamma(\theta)$ with $0 \leq \alpha(c) < \alpha(h) \leq 1$ (so that honest types are more likely to succeed than corrupt ones) and $\gamma(\cdot)$ increasing, possibly weakly, with $0 < \gamma(\underline{\theta}) \leq \gamma(\overline{\theta}) < 1$.

Not all voters may directly observe the project outcome. Specifically, a villager observes $\omega_v^t \in \{0, 1\}$ with probability $\lambda \in (0, 1]$. With the complement probability $1 - \lambda$, the voter observes nothing. Notice that, since we assume a mass of voters, the above implies that a proportion λ of villagers are informed about ω_v^t.

When casting their ballot at the end of the first period, voters behave prospectively. Each voter cares about project completion in her village, and therefore has a preference for the candidate who is more likely to deliver a success in the second period, that is, more likely to be an honest type. In addition, her evaluation of politicians also depends on an idiosyncratic shock, denoted σ_i^J, with $J \in \{I, C\}$. It is common knowledge that $\sigma_i^I - \sigma_i^C$ is continuously distributed according to the CDF $G(\cdot)$ over the interval $[-1, 1]$, with $G(-1) > 0$ and $G(1) < 1$ (these two inequalities just guarantee vote shares are always interior). In short, each voter's second-period utility function can be represented as:

$$u_i(\omega_v^2) = \omega_v^2 + \mathbb{I}_I \sigma_i^I + (1 - \mathbb{I}_I)\sigma_i^C, \tag{1}$$

with $\mathbb{I}_I$ an indicator function equal to 1 if the incumbent I is re-elected. At time of elections, villagers vote sincerely for the candidate who maximizes their expected second period utility.

Informational campaigns

In our setting, voters are relatively poorly informed. Villagers do not know their incumbent's type (though they have some prior about it). They do not necessarily observe whether the public good project has been successful in their village. And they do not know the economic circumstances in which office-holders operate (i.e., the theta). This approach is by design since we are interested in the effect of informational campaigns on electoral accountability.

We define an informational campaign as the random selection of villages by researchers, who then randomly assign a subset of the sample to the treatment, with the others forming the control group. We consider two kinds of interventions, distinguished by the nature of the treatment. In the first type of intervention, *all* voters in treated villages are informed about the outcome of the project in their village. We label this form of intervention *performance treatment.* In the second type of intervention, all voters in treated villages are informed about their representative's type. Since an incumbent is either honest or corrupt, we refer to this form of intervention as *corruption treatment.*

Timing

The game, in turn, proceeds as follows

Period 1:

1. Nature draws politicians' types (incumbents' and challengers') in all districts;
2. Nature draws the shock θ^1, and the first-period project outcome is realized in each village;
3. A voter learns the project outcome in her village v with probability λ. Some villages are randomly selected for intervention, with a subset of them being treated.
4. Each voter i observes the valence shocks σ_i^I and σ_i^C and then casts a vote for I_d or C_d. In each district, the incumbent I_d is re-elected or replaced by C_d.

Period 2:

1. Nature draws the second-period shock and second-period project outcome realized in each village;
2. The game ends and payoffs are realized.

The equilibrium concept is Perfect Bayesian Equilibrium.

Remarks on the set-up

In our model, we deal away with any moral hazard concerns. There are several reasons for this. First, many information campaigns are purposefully small and inconsequential for the aggregated electoral result so as not to affect incumbents' behavior (though, we note that candidates or their agents do sometimes react to the campaign, see Arias et al., 2020; Sircar and Chauchard, 2020). Eliminating moral hazard from the model is a way to guarantee this assumption holds by definition, i.e., it is an ideal scenario. Second, the absence of strategic choices by the incumbents significantly simplify the exposition of the analysis and allows us to focus on the empirical implications of our results. Relatedly, focusing on voters' selection problem allows us to perfectly isolate the effect of additional information. From the onset, let us stress that versions of all our results hold when office-holders take strategic action and when they can react to the intervention.

Our set-up is flexible enough to encompass many different empirical settings. It can be adapted to the study of national legislators ($D > 1$, $V \geq 1$), or village heads ($D > 1$, $V = 1$). Our framework also allows to think about different sampling frames. Researchers can intervene in villages drawn all over the country (country-wide analysis) or drawn within a single district (within district analysis). In Section 7, we also discuss how our model can be used to think about analyzing randomization at the level of individuals, when in a village, some voters are treated and others belong to the control group.

Regarding villager information, we assume our voters are imperfectly informed (i) about their incumbent's type, (ii) about the state of the world θ^1, and (iii) about their representative's performance in their village. The first and second points are relatively common in formal models (for two distinct frameworks which include both assumptions in different forms, see Canes-Wrone et al., 2001; Besley, 2006). The most controversial choice regards point (iii) (though some assume the existence of noise voters, e.g., Calvert, 1985). In our cases, we include completely uninformed voters so that performance treatments can have an effect. Let us stress that none of our results are affected if all villagers observe the outcome of the public good project in their village.

Finally, let us introduce some useful terminology, especially for our discussion of empirical models. We distinguish between what we call *context* and what we refer to as *circumstances*. The context corresponds to the primitives of the model. The political context is the distribution of types among politicians (π^I and π^C). The economic context consists of the distribution of economic shocks ($F(\cdot)$ over $[\underline{\theta}, \overline{\theta}]$). The local context is the function transforming office-holder's type and economic shock into project outcome $Pr(\omega_v^1 = 1) = \alpha(\tau)g(\theta)$, etc. The circumstances, in turn, correspond to the actual realization of all random variables. The political circumstance corresponds to the actual type of the office-holder ($\tau \in \{c, h\}$). The economic context is the actual value of the economic shock (θ). The local context is the actual realization (failure or success) of the public good project

($\omega_v^1 \in \{0,1\}$), etc. In all that follows, we keep the context fixed (the fundamentals are always the same), but we allow the circumstances to vary (the realizations of variable is not held constant).

4 Analysis

As our framework focuses on the effect of information, the analysis of the model centers around villagers' beliefs. All voters share the same prior belief, therefore posteriors vary solely according to the information received, either by Nature or via an information treatment by researchers. We denote $\mu_v(z, \iota_v)$ the posterior of a voter from village v that the incumbent is honest. The first argument captures the information that the voter receives from Nature, independent of any information treatment. Recall that, absent an information campaign, voters either learn nothing or observe the outcome of the project in their village. Thus, $z = \{\omega_v^1, \emptyset\}$. The second argument instead captures the information (if any) that the voter receives from researchers. Recall that we are considering two types of information treatments: performance (all voters in treated villages learn about project outcome ω_v^1) and corruption (all voters in treated villages learn about incumbent's type τ_{I_d}). Thus, $\iota_v \in \{\emptyset, \omega_v^1, \tau_{I_d}\}$, where $\iota_v = \emptyset$ indicates that the voter lives in a village that receives no information treatment. Combining the information that may be disclosed by Nature and by researchers, we obtain six possibilities. These are summarized in Table 1.

		From Nature	
		No info	Project outcome
From Researchers	No intervention	$\mu_v(\emptyset,\emptyset)$	$\mu_v(\omega_v^1,\emptyset)$
	Performance	$\mu_v(\emptyset,\omega_v^1)$	$\mu_v(\omega_v^1,\omega_v^1)$
	Corruption	$\mu_v(\emptyset,\tau_{I_d})$	$\mu_v(\omega_v^1,\tau_{I_d})$

Table 1: Posteriors as a function of information

Quite naturally, if the voter learns nothing from Nature or researchers, her posterior equals her prior: $\mu_v(\emptyset,\emptyset) = \pi^I$. In turn, whether the voter learns the project outcome via Nature, informal campaign, or both, her posterior is always the same. By Bayes' rule, $\mu_v(1,\emptyset) = \mu_v(\emptyset,1) = \mu_v(1,1) = \frac{\pi^I\alpha(h)}{\pi^I\alpha(h)+(1-\pi^I)\alpha(c)}$ after completed project and $\mu_v(0,\emptyset) = \mu_v(\emptyset,0) = \mu_v(0,0) = \frac{\pi^I(1-\gamma^e\alpha(h))}{\pi^I(1-\gamma^e\alpha(h))+(1-\pi^I)(1-\gamma^e\alpha(c))}$ after uncompleted project, with γ^e the expected contribution of the economic environment to the success of a project in the voter's village, formally $\gamma^e = \int_{\underline{\theta}}^{\overline{\theta}} \gamma(\theta) dF(\theta)$. Finally, the corruption treatment fully reveals the incumbent's type so $\mu_v(z, h) = 1$ and $\mu_v(z, c) = 0$ for all $z \in \{\omega_v^1, \emptyset\}$.

Turning now to villagers' electoral decision, a voter's expected payoff from voting for the incumbent as a function of her information (z, ι_v) is: $\big(\mu_v(z, \iota_v)\alpha(h) + (1 - \mu_v(z, \iota_v))\alpha(c)\big) \times \gamma^e + \sigma_i^I$. A voter expects a successful project with probability $\alpha(h)\gamma^e$ in period 2 if the incumbent is honest and expects a completed project with probability $\alpha(c)\gamma^e$ if the incumbent is corrupt. In addition, she gets an additional payoff σ_i^I from voting for the incumbent. In turn, because the voter has no information about the challenger, she can only rely on her prior, and her expected payoff from casting a ballot for C_d is: $\big(\pi^C\alpha(h) + (1 - \pi^C)\alpha(c)\big) \times \gamma^e + \sigma_i^C$. Rearranging a bit, we obtain that the voter votes for the current office-holder if and only if (ties being a zero probability event):

$$\sigma_i^I - \sigma_i^C \geq -(\mu_v(z, \iota_v) - \pi^C)(\alpha(h) - \alpha(c))\gamma^e \tag{2}$$

From this, we can easily obtain vote shares in a village v taking advantage that (i) we assume throughout a unit mass of voters in each village (with λ of them informed) and (ii) the difference $\sigma_i^I - \sigma_i^C$ is continuously distributed. The realized incumbent's vote share in village v is then:

$$\begin{aligned} S_v(\omega_v^1, \iota_v) := &\lambda \left(1 - G\big((\pi^C - \mu_v(\omega_v^1, \iota_v))(\alpha(h) - \alpha(c))\gamma^e\big)\right) \\ &+ (1 - \lambda) \left(1 - G\big((\pi^C - \mu_v(\emptyset, \iota_v))(\alpha(h) - \alpha(c))\gamma^e\big)\right). \end{aligned} \tag{3}$$

Noting that a voter's posterior upon learning the project outcome is the same whether the information comes from Nature or from an intervention, we can easily compute vote shares in villages under all possible treatment and control conditions (the proof of this Lemma is omitted as it follows directly from Equation 3).

Lemma 1. *The incumbent's vote share is*

(i) $S_v(\omega_v^1, \emptyset) = \lambda\Big(1 - G\big((\pi^C - \mu_v(\omega_v^1, \emptyset))(\alpha(h) - \alpha(c))\gamma^e\big)\Big) + (1 - \lambda)\Big(1 - G\big((\pi^C - \pi^I)(\alpha(h) - \alpha(c))\gamma^e\big)\Big)$ *in a non-treated village;*

(ii) $S_v(\omega_v^1, \omega_v^1) = 1 - G\big((\pi^C - \mu_v(\omega_v^1, \omega_v^1))(\alpha(h) - \alpha(c))\gamma^e\big)$ *in a village treated with the performance treatment;*

(iii) $S_v(\omega_v^1, \tau_{I_d}) = 1 - G\big((\pi^C - \mu_v(\omega_v^1, \tau_{I_d}))(\alpha(h) - \alpha(c))\gamma^e\big)$ *in a village treated with the corruption treatment.*

Lemma 1 highlights an important substantive difference between performance and corruption treatments. Performance treatments extend the reach of existing information, the outcome of the village-specific public good project, which only reach a proportion $\lambda \in (0, 1]$ of villagers absent intervention (recall that $\mu_v(\omega_v^1, \emptyset) = \mu_v(\omega_v^1, \omega_v^1)$). Note that voters who are informed by Nature absent treatment can be understood as non-compliers from the point of view of researchers. In contrast, corruption treatments provide information that is never accessible to the villagers without the researchers' intervention. This difference has, we will show below, subtle but heavy consequences for the accumulation of knowledge on the topic. To describe these repercussions, we now study how our theoretical framework helps uncovering some theoretical properties of empirical estimates.

5 Theoretical implications of empirical models

Before proceeding to the analysis of the theoretical implications of empirical models, it is useful to (re)define a few terms we will use in what follows. Uncontroversially, *a research design is unbiased* if it does not introduce bias in the estimate of a particular estimand. In turn, we say that *an empirical specification permits comparability* if, fixing the context (i.e., the fundamental underlying parameters), the regression researchers run yields an estimate of the same estimand for all possible draws of observations. Put differently, comparability requires specifications to refer to the same counterfactual. Finally, we say that *two studies are comparable* if they use an empirical specification that permits comparability and *two studies permit cumulative knowledge* if they are comparable and use an unbiased research design.

It is important to stress that we do not require external validity for studies to be comparable. Indeed, the two notions are very much distinct. Comparability depends on the empirical specification; external validity is a property of the estimand. However, it is apparent that comparability is a necessary condition to assess the external validity of a study's findings. If studies are not comparable, then researchers can never distinguish whether the varying estimates they obtain are the result of a lack of external validity or are due to their empirical specification measuring different quantities.

In this paper, we are solely concerned with comparability. As noted above, we suppose that any intervention randomly select a set of villages and randomly assigned a treatment to a subset of the sample. There is no source of bias, and studies permit cumulative knowledge if and only if they are comparable. Further, we eliminate any external source of variation by fixing the context throughout (the distribution of types, the distribution of economic shocks, the production technology of local public good, the probability villagers are informed by Nature). We can then think of each study consisting of a draw of observations at different points in time and/or from distinct locations in a given context. As such, each study analyzes the effect of interventions in different circumstances (different realizations of the random variables).

But how should we study the electoral consequences of informational campaign? In this section, we assess whether the empirical specification commonly used in the literature to evaluate the impact of information campaigns permits comparability. We follow Dunning et al. (2020), Ferraz and Finan (2004), Larreguy et al. (2020) among others, in analyzing the effect of information treatments conditional on the information provided being 'good news' or 'bad news.' We label this empirical approach the *conditional differences in mean vote shares*.

As in the literature, good news are defined as pieces of information that are expected to improve voters' evaluations of the incumbent. Similarly, bad news, in expectation, lower the voters' posterior on the incumbent's relevant characteristics. While good or bad news may be hard to operationalize in practice, in our set-up they have a natural definition. When it comes to the incumbent's performance, project completion constitutes good news, whereas failure denotes bad news. For corruption treatments, treated voters receive good news whenever they are informed that their incumbent is an honest type and bad news whenever he is revealed to be corrupt.

From an empirical standpoint, denote Y_{vds} the incumbent's vote share in village v in district d in state s, $\Omega_{vds} \in \{0,1\}$ the project's outcome ($\Omega_{vds} = 1$ indicating success), $H_{vds} \in \{0,1\}$ the incumbent's type ($H_{vds} = 1$ indicating honest), and $T_{vds} \in \{0,1\}$ whether a village v is treated ($T_{vds} = 1$ indicating treatment). The conditional difference in means vote share is simply: $E(Y_{vds}|T_{vds} = 1, \Omega_{vds} = \Omega) - E(Y_{vds}|T_{vds} = 0, \Omega_{vds} = \Omega)$ for the performance treatment (with $\Omega = 1$ for good news and $\Omega = 0$ for bad news) and $E(Y_{vds}|T_{vds} = 1, H_{vds} = H) - E(Y_{vds}|T_{vds} = 0, H_{vds} = H)$ for the corruption treatment ($H = 1$ for good news, $H = 0$ for bad news). In practice, this is equivalent to running the following interaction models:

$$Y_{vds} = a_0 + a_1\Omega_{vds} + a_2T_{vds} + a_3\Omega_{vds} \times T_{vds} + \epsilon_{vds} \quad \text{for the performance treatment;} \tag{4}$$

$$Y_{vds} = b_0 + b_1H_{vds} + b_2T_{vds} + b_3T_{vds} \times H_{vds} + \varepsilon_{vds} \quad \text{for the corruption treatment.} \tag{5}$$

In these models, the conditional difference in mean vote shares for bad news is a_2 for the performance treatment and b_2 for the corruption treatment. The conditional difference for good news equals $a_2 + a_3$ for the performance treatment and $b_2 + b_3$ for the corruption treatment.

We now use our model to map these empirical estimates into their equivalent theoretical quantities. This allows us to study their properties. In doing so we assume that researchers draw a random mass of villages, so as to eliminate any source of statistical noise.

First, we ask what is the expected effect (if any) of informational campaigns on electoral outcomes.

Proposition 1. *For the performance treatment, bad news strictly reduces the incumbent's vote share and good news strictly increases it unless $\lambda = 1$.*
For the corruption treatment, bad news strictly reduces the incumbent's vote share and good news strictly increases it.

The treatment has an effect as long as the information provided is new to some voters in treated villages. For performance treatments, this requires that some villagers do not learn the outcome of their local public good project in the absence of an intervention, that is $\lambda < 1$. If so, good news improves the electoral fortune of the office-holder. Since we assume $\alpha(h) > \alpha(c)$, honest types are more likely to be successful than corrupt politicians. Therefore, project completion is a signal that raises voters' evaluation of the incumbent, project failure is a signal that depresses villagers' opinion of their office-holder.

As noted above, the information provided via corruption treatments is never available to villagers absent intervention. In some cases, however, observing project outcomes is a perfect substitute for directly learning the incumbent's type. This occurs when honest types always succeed and corrupt office-holders always fail. This, however, is an extreme case, which we exclude with our assumptions on the effect of the economic circumstances on project realization ($0 < \gamma(\underline{\theta}) \leq \gamma(\overline{\theta}) < 1$ so honest incumbents fail with positive probability). Consequently, the corruption treatments always have an effect in our set-up.

In what follows, to avoid dealing with too many cases, but without much loss of generality, we will assume that $\lambda < 1$ so not all villagers learn the outcome of their local public good project absent intervention. We also impose $\alpha(c) > 0$ so corrupt types produce some successful outcomes. The theory then predicts that the effect of information treatments conforms with intuition: bad news reduces the incumbent's vote share (so $a_2 < 0$ and $b_2 < 0$), and good news improves it (so $a_2 + a_3 > 0$ and $b_2 + b_3 > 0$).

With this, we are now ready to answer our main question: Does the most commonly used conditional difference in mean vote shares permit comparability? The next proposition indicates that studies that employ this empirical specification are always comparable when it comes to the performance treatment. When it comes to the corruption treatment, comparability is by no mean guaranteed. It holds if and only if economic circumstances have no effect on performance.

Proposition 2. *The conditional difference in mean vote shares permits comparability if the treatment is the performance treatment.*
When it comes to the corruption treatment, the conditional difference in mean vote shares permits comparability if and only if $\gamma(\underline{\theta}) = \gamma(\overline{\theta})$.

Let us provide some intuition for this result. Using Lemma 1, we can precisely map the coefficients in Equation 4 and Equation 5 into our theoretical framework. For performance treatment, the effect of bad news (a_2) and good news ($a_2 + a_3$) are, respectively:

$$\begin{aligned} a_2 =& E\big(S_v(\omega_v^1, \iota_v)|\omega_v^1 = 0, \iota_v = \omega_v^1\big) - E\big(S_v(\omega_v^1, \iota_v)|\omega_v^1 = 0, \iota_v = \emptyset\big) \\ =& (1-\lambda)\Big(G\big((\pi^C - \pi^I)(\alpha(h) - \alpha(c))\gamma^e\big) - G\big((\pi^C - \mu^v(0,0))(\alpha(h) - \alpha(c))\gamma^e\big)\Big) \end{aligned} \tag{6}$$

$$\begin{aligned} \text{and} \quad a_2 + a_3 =& E\big(S_v(\omega_v^1, \iota_v)|\omega_v^1 = 1, \iota_v = \omega_v^1\big) - E\big(S_v(\omega_v^1, \iota_v)|\omega_v^1 = 1, \iota_v = \emptyset\big) \\ =& (1-\lambda)\Big(G\big((\pi^C - \pi^I)(\alpha(h) - \alpha(c))\gamma^e\big) - G\big((\pi^C - \mu^v(1,1))(\alpha(h) - \alpha(c))\gamma^e\big)\Big) \end{aligned} \tag{7}$$

Equation 6 and Equation 7 indicate that the circumstances in which the study took place play no role. The effect of the performance treatment (whether it delivers good or bad news) is only a function of the context the researchers intervene in—the distributions of honest and corrupt types, the proportion of informed voters in the control group, the distribution of economic shocks, and the production technology of public good. Since all fundamentals are assumed to be fixed, these parameters remain constant from one draw of observations to the next. As a consequence, all studies drawing observations from the same context and running regression Equation 4 estimate the same estimand: the (ITT) effect of performance treatment on incumbents' electoral performance. Hence, all such studies are comparable.

Let us now turn to the corruption treatment. Using Equation 5, the conditional difference in means compares treated and control villages who all are represented by a corrupt type for bad news or by a honest type for good news. With the help of Lemma 1, we can again relate the coefficient in Equation 5 to their theoretical equivalents.

$$
\begin{aligned}
b_2 =& E_{\omega_v^1}\big(S_v(\omega_v^1, \iota_v)|\tau_{I_d} = c, \iota_v = \tau_{I_d}\big) - E_{\omega_v^1}\big(S_v(\omega_v^1, \iota_v)|\tau_{I_d} = c, \iota_v = \emptyset\big) \\
=& \Big(1 - G\big(\pi^C(\alpha(h) - \alpha(c))\gamma^e\big)\Big) \\
& - \lambda\alpha(h)\gamma(\theta)\Big(1 - G\big((\pi^C - \mu_v(1, \emptyset))(\alpha(h) - \alpha(c))\gamma^e\big)\Big) \\
& - \lambda\big(1 - \alpha(h)\gamma(\theta)\big)\Big(1 - G\big((\pi^C - \mu_v(0, \emptyset))(\alpha(h) - \alpha(c))\gamma^e\big)\Big) \\
& - (1 - \lambda)\Big(1 - G\big((\pi^C - \pi^I)(\alpha(h) - \alpha(c))\gamma^e\big)\Big) \\
=& (1 - \lambda)\Big(G\big((\pi^C - \pi^I)(\alpha(h) - \alpha(c))\gamma^e\big) - G\big(\pi^C(\alpha(h) - \alpha(c))\gamma^e\big)\Big) \\
& + \lambda\alpha(c)\gamma(\theta)\Big(G\big((\pi^C - \mu_v(1, \emptyset))(\alpha(h) - \alpha(c))\gamma^e\big) - G\big(\pi^C(\alpha(h) - \alpha(c))\gamma^e\big)\Big) \\
& + \lambda\big(1 - \alpha(c)\gamma(\theta)\big)\Big(G\big((\pi^C - \mu_v(0, \emptyset))(\alpha(h) - \alpha(c))\gamma^e\big) - G\big(\pi^C(\alpha(h) - \alpha(c))\gamma^e\big)\Big)
\end{aligned} \tag{8}
$$

and

$$
\begin{aligned}
b_2 + b_3 =& E_{\omega_v^1}\big(S_v(\omega_v^1, \iota_v)|\tau_{I_d} = h, \iota_v = \tau_{I_d}\big) - E_{\omega_v^1}\big(S_v(\omega_v^1, \iota_v)|\tau_{I_d} = h, \iota_v = \emptyset\big) \\
=& (1 - \lambda)\Big(G\big((\pi^C - \pi^I)(\alpha(h) - \alpha(c))\gamma^e\big) - G\big((\pi^C - 1)(\alpha(h) - \alpha(c))\gamma^e\big)\Big) \\
& + \lambda\alpha(h)\gamma(\theta)\Big(G\big((\pi^C - \mu_v(1, \emptyset))(\alpha(h) - \alpha(c))\gamma^e\big) - G\big((\pi^C - 1)(\alpha(h) - \alpha(c))\gamma^e\big)\Big) \\
& + \lambda\big(1 - \alpha(h)\gamma(\theta)\big)\Big(G\big((\pi^C - \mu_v(0, \emptyset))(\alpha(h) - \alpha(c))\gamma^e\big) - G\big((\pi^C - 1)(\alpha(h) - \alpha(c))\gamma^e\big)\Big)
\end{aligned} \tag{9}
$$

Equation 8 and Equation 9 highlight that the conditional difference in mean vote shares is now the weighted average of three components. The first component is the effect of the treatment for voters that do not observe the outcome of the project in their village (this is captured in the first line of both equations). The second component is the impact of revealing the incumbent's type to villagers that observe project success (in the second line). The third is the effect of the treatment for voters that observe a failure (in the third line).

Observe that all three components only depend on the context in which the studies take place. As for performance treatments, they are only a function of the distributions of politicians' types, the distribution of economic shocks, etc. But the coefficients above are not a simple sum of these three components, they are each a weighted average. And the weights are a function of the economic *circumstances* in which researchers intervene. Indeed, when economic circumstances are favourable (θ is high), relatively many villages experience successful outcomes and so the obtained coefficients puts a high weight on the impact of revealing the incumbent's type in villagers with completed public goods. In turn, when economic circumstances are unfavourable (θ is low), relatively few villages see local public good project comes to fruition, and the coefficients puts significant weight on the effect of the treatment for voters that observe a failure.

Consequently, unless the economic circumstances have no effect on performance ($\gamma(\underline{\theta}) = \gamma(\overline{\theta})$), two studies which intervene in the same exact context yield estimates of different, *circumstance-specific*, estimands. When $\gamma(\theta)$ is not constant, they measure the effect of the corruption treatment in the economic circumstances at the time of their intervention. In short, the proportion of villagers experiencing a failed versus a successful project is balanced between treated and control villages *within* a study, thanks to randomization and the draw of a mass of observations. However, the proportion of villages in the treated and in the control group which experience successful public good projects is not balanced *across* studies. And this generates problems for comparability.

In concluding this section, it is important to stress that the problem we identify is not specific to corruption treatments. Indeed, it arises whenever two conditions are met. First, researchers provide information that is never accessible to voters absent an intervention (in our case this is the incumbent's type). And second, voters who reside in control villages nonetheless observe other pieces of information that allow them to update their beliefs about the incumbent (in our case, this is the project outcome). When these conditions are met, the total effect of the treatment depends on the exact composition of the control group (as in Equation 8 and Equation 9) which (is very likely to) be specific for the draw of observations obtained by the researchers. As above, studies then only measure *circumstance-specific estimands*, and comparability is not achieved.

6 Possible remedies for comparability

Having exposed conditions under which the conditional difference in mean vote shares permits or not comparability, we now turn to possible remedies.

The comparability issue we highlight arises when the proportion of voters who have relatively good opinion of the incumbent independent of researchers' intervention changes from one study to the next (even fixing the context). One may think that Equation 5 allows to test for this possibility. After all, if villagers never observe the project outcome ($\lambda = 0$) or learn nothing about the incumbent's characteristics from success or failure ($\alpha(c) = \alpha(h)$), then the incumbent's vote-share in control villages depends solely on the voters' prior beliefs. It is uncorrelated with incumbent's type. This would imply that, in Equation 5, the estimated coefficient $\hat{b}_1$ should be close to zero and fail to reach statistical significance. Unfortunately, this is not the unique possible interpretation of a non-statistically significant $\hat{b}_1$. Another competing reason is simply that the treatment is orthogonal to the information voters use to make their electoral decision. In other words, the intervention does not provide relevant information to voters and any effect obtained (positive or negative $\hat{b}_2$ and $\hat{b}_3$) could be due to statistical noise. Hence, this first solution puts researchers in a quandary. It could be that the conditional difference in means permits comparability in the precise context they intervene in, or it may be that their intervention is ineffective.

In what follows, we argue for a more radical departure from the usual conditional difference in mean vote shares. Our suggestion involves a change of counterfactuals. Instead of asking what if the control group were to be treated (as researchers implicitly do when employing Equation 5) we propose that researchers ask: what if the treated units had not received treatment? Instead of conditioning on what control units would have observed if treated, researchers should condition on what the treated villages would have learned absent treatment. We label this approach the *augmented conditional difference in mean vote shares*, augmented because researchers condition on what the control group would have observed if treated *and* on what treated units would have learned absent treatment.

In practice, this corresponds to running the following saturated model:

$$Y_{vd} = c_0 + c_1 H_{vd} + c_2 \Omega_{vd} + c_3 T_{vd} + c_4 T_{vd} \times H_{vd} + c_5 T_{vd} \times \Omega_{vd} + c_6 H_{vd} \times \Omega_{vd} + c_7 T_{vd} \times H_{vd} \times \Omega_{vd} + \varepsilon_{vd} \quad (10)$$

The next proposition highlights the main advantage of the augmented conditional difference in mean vote shares, namely that it permits comparability.

Proposition 3. *The augmented conditional difference in mean vote shares always permits comparability for corruption treatments.*

For all samples drawn from a given context, the augmented conditional difference in mean vote shares estimates the same estimands. This, however, comes at a cost. Researchers recover not one, but multiple effects for bad and good news. Indeed, fixing the type of news, there will be one estimand for villages where the project failed (c_3 for bad news, $c_3 + c_4$ for good news) and one for those where the project was completed ($c_3 + c_5$ for bad news and $c_3 + c_4 + c_5 + c_7$ for good news). This means that researchers need to go beyond the simple dichotomy bad-good news.

Why does the augmented conditional difference permits comparability when the simple conditional difference fails? We can again use Lemma 1 to recover the theoretical equivalent to the empirical effect of bad news, c_3 and $c_3 + c_5$ (a similar reasoning holds for good news). After slight rearranging, we obtain:

$$\begin{aligned}
c_3 =& E\big(S_v(\omega_v^1, \iota_v) | \tau_{I_d} = c, \omega_v^1 = 0, \iota_v = \tau_{I_d}\big) - E\big(S_v(\omega_v^1, \iota_v) | \tau_{I_d} = c, \omega_v^1 = 0, \iota_v = \emptyset\big) \\
=& (1-\lambda)\Big(G\big((\pi^C - \pi^I)(\alpha(h) - \alpha(c))\gamma^e\big) - G\big(\pi^C(\alpha(h) - \alpha(c))\gamma^e\big)\Big) \\
&+ \lambda\Big(G\big((\pi^C - \mu_v(0, \emptyset))(\alpha(h) - \alpha(c))\gamma^e\big) - G\big(\pi^C(\alpha(h) - \alpha(c))\gamma^e\big)\Big) \qquad (11) \\
\text{and} \quad c_3 + c_4 =& E\big(S_v(\omega_v^1, \iota_v) | \tau_{I_d} = c, \omega_v^1 = 1, \iota_v = \tau_{I_d}\big) - E\big(S_v(\omega_v^1, \iota_v) | \tau_{I_d} = c, \omega_v^1 = 1, \iota_v = \emptyset\big) \\
=& (1-\lambda)\Big(G\big((\pi^C - \pi^I)(\alpha(h) - \alpha(c))\gamma^e\big) - G\big(\pi^C(\alpha(h) - \alpha(c))\gamma^e\big)\Big) \\
&+ \lambda\Big(G\big((\pi^C - \mu_v(1, \emptyset))(\alpha(h) - \alpha(c))\gamma^e\big) - G\big(\pi^C(\alpha(h) - \alpha(c))\gamma^e\big)\Big) \qquad (12)
\end{aligned}$$

The impact of the corruption treatment is now the sum of two effects: the effect on uninformed voters (first line in Equation 11 and Equation 12) and the effect on villagers who learn the outcome of the public good project in their village independently of the researchers' intervention (second line in Equation 11 and Equation 12). All voters informed by Nature now receive the same information—project failure in Equation 11, project success in Equation 12—, and the specific economic circumstances do not play any role any more. There is no longer an issue of sample imbalance *across studies* as the coefficients are only a function of the context in which the studies take place. That is, all studies estimate the same estimands.

What are these estimands? The coefficient c_3 corresponds to the effect of information in villages where project failed and the sum of c_3 and c_4 captures the effect of information in villages where the project was successful. Both can be understood as measuring the *pure* effect of informational campaign. Both are related to clear theoretical quantities of interest—what happens when bad news is provided after good/bad office-holder's performance—, rather than relative to a moving baseline—what happens when bad news is provided in particular economic circumstances.

The theoretical quantities recovered in Equation 10 are of policy relevance even if they do not match any specific situation. The coefficient c_3 can be interpreted as the impact of bad news when all villages experience unsuccessful projects, $c_3 + c_4$ its electoral consequences when all villages see successful projects. In all informational campaigns, the proportion of successful projects will fall between these two extremes so that the effect will be in-between the two coefficients. In other words, the minimum of c_3 and $c_3 + c_4$ corresponds to the greatest possible decline for the incumbent's vote share following bad news (recall $c_3 < 0$ and $c_3 + c_4 < 0$), the maximum of c_3 and $c_3 + c_4$ measures the lowest possible drop in the incumbent's electoral results: $\min\{c_3, c_3 + c_4\} \leq b_2 \leq \max\{c_3, c_3 + c_4\}$, where b_2 is the coefficient obtained from the usual conditional differences in mean vote shares (Equation 5). Similarly, the sums of coefficients $c_3 + c_5$ and $c_3 + c_5 + c_7$ bound the electoral consequences of good news: $\min\{c_3 + c_5, c_3 + c_5 + c_7\} \leq b_3 + b_4 \leq \max\{c_3 + c_5, c_3 + c_5 + c_7\}$. (In the special case when bad news perfectly reveals the incumbent is corrupt and good news perfectly reveals he is honest, $c_3 + c_4 < c_3$ and $c_3 + c_5 + c_7 < c_3 + c_5$, but these inequalities do not hold for all interventions revealing novel information to voters.) Hence, our approach can provide a way to assess whether the benefit of informational campaigns is worth the cost of such interventions.

Still, the solution we propose is no panacea. We recognize two difficulties associated with the augmented conditional difference in means vote shares, one practical and one theoretical. In practice, the empirical specification we suggests require that researchers have a deep knowledge about the local circumstances, and the funds available to collect a large amount of data prior to randomization. In theory, our augmented conditional difference permits comparability only if researchers are able to condition on *all electorally relevant information* available to the villagers absent the intervention. Such variables may be hard to identify or even observe for the researchers. For example, some relevant village-level variables may not be measurable or pieces of information that change voters' evaluation of incumbents may not be detectable. If scholars cannot condition on all villagers' information absent treatment, the estimands again becomes a function of local circumstances. The problem we documented in the commonly used conditional differences in mean vote shares then also plagues the augmented conditional difference (Appendix A highlights this problem using the potential outcome framework).

Overall, until now, our results offer more bad news than good news for cumulative knowledge. The conditional difference in mean vote shares used in most studies (where researchers only condition on what the control would have observed if treated) is likely to prevent comparability. The augmented conditional difference provides a way forward only if all information observed by villagers absent intervention is conditioned on (i.e., only if the treated units perfectly resemble the control units absent treatment). When this requirement is not met, scholars should doubt that studies achieve comparability. Any difference in estimates across settings, even after fixing the information treatment, can be attributed to low external validity (contexts are too different) or to low comparability (each study measuring a different estimand even after fixing the context). It becomes questionable, if not impossible, to ascertain cumulative knowledge.

So far, we have considered studies that focus on aggregate outcomes: the incumbent's vote-share at the village level. In the following section we will discuss one potential remedy for comparability that can be adopted when researchers instead consider individual-level outcomes.

7 Individual-level outcomes

For reasons of cost, researchers often do not have the ability to select a sample of villages and divide it between treated and control units (e.g., Ferraz and Finan, 2004, takes advantage of a government program to analyze municipality-level outcomes). Rather, researchers intervene in a few villages where, in each, they survey some voters and provide a subset of respondents with additional information (e.g., Dunning et al., 2020). Scholars then look at individual-level outcomes. They analyze the effect of the informational campaign on individuals' vote choice (or voting intentions) using an empirical specification which typically includes village fixed effect so that any village specific attributes is removed from the comparison between treated and control units.

More specifically, denote Y_{ivds} the reported voting intention (or vote choice) for the incumbent by individual i in village v in district d, $T_{ivd} \in \{0,1\}$ whether an individual i is treated in village v, and again $\Omega_{vd} \in \{0,1\}$, $H_{vd} \in \{0,1\}$ the information provided ($\Omega = 1/H = 1$ denoting successful project/honest incumbent as before). The conditional difference in mean vote choices (or intentions) then becomes $E(Y_{ivd}|T_{vd} = 1, \Omega_{vd} = \Omega) - E(Y_{ivd}|T_{vd} = 0, \Omega_{vd} = \Omega)$ for performance treatment ($\Omega \in \{0,1\}$, $E(Y_{ivd}|T_{vd} = 1, H_{vd} = H) - E(Y_{ivd}|T_{vd} = 0, H_{vd} = H)$ for the corruption treatment ($H \in \{0,1\}$). In term of regressions, this is akin to:

$$Y_{ivd} = \delta_v + d_0 + d_1\Omega_{vd} + d_2T_{ivd} + d_3\Omega_{vd} \times T_{ivd} + \epsilon_{vd} \quad \text{for performance treatment;} \tag{13}$$

$$Y_{ivd} = \delta_v + e_0 + e_1H_{vd} + e_2T_{ivd} + e_3T_{ivd} \times H_{vd} + \varepsilon_{vd} \quad \text{for corruption treatment,} \tag{14}$$

with δ_v a village fixed effect.

There is nothing special about individual-level outcomes. For performance treatments (Equation 13), researchers still recover an intention to treat effect of increasing the reach of existing information (d_2 for bad news, project failure, and $d_2 + d_3$ good news, project success). For corruption treatment (Equation 14), regression coefficients yield the average treatment effect of providing new information for bad news (incumbent is corrupt, e_2) and good news (incumbent is honest, $e_2 + e_3$).

As a result, somewhat unsurprisingly, the conditional difference in means for corruption treatments does not permit comparability in our set-up. Within in a given village, the difference in voting intention between treated and control villagers is a function of the project outcome (e.g., good news have less of an impact when the local public good has been completed). Across all villages in the sample, the estimate of the electoral consequences of the corruption treatment is a weighted average of the effect in villages which experience successful projects and the effect in villages which see unsuccessful projects. If the sample is representative of all villages, then the proportion of villages that experience success in the sample is $\alpha(\tau)\gamma(\theta)$. Thus, the weights in the weighted average are again a function of the circumstances in which the survey take place. And comparability is not achieved.

Proposition 4. *Suppose the randomization is at the individual level and the outcome of interest is individual voting intention.*
The conditional difference in mean voting intentions permits comparability if the treatment is the performance treatment.
When it comes to the corruption treatment, the conditional difference in mean voting intentions permits comparability if and only if $\gamma(\underline{\theta}) = \gamma(\overline{\theta})$.

We can then naturally extend the result to the augmented conditional difference in mean vote intentions. As before, this empirical specification can offer a potential solution to ensure comparability. However, comparability requires that researchers account for all possible electorally relevant pieces of information villagers may receive in the control condition, an hard-to-achieve feat for researchers.

So far, it looks like more of the same rather than a way out of the conundrum of comparability, or lack thereof. But researchers can exploit another avenue when implementing individual-level randomization. They can make use of a well-known property of voters' evaluation of their representative. This prior opinion (prior because measured before applying the treatment) captures all the relevant information available to villagers, without the need for researchers to make any assumptions. In other words, villagers' interim belief (to use formal language) is a sufficient statistic for all the factors that affect voters' view of their office-holder absent intervention. Conditioning on them, hence, resolve all problems of balance across studies and allows for comparability. Further, information on voters' prior evaluations is relatively easy for researchers to collect. This can be done, for example, via the use of a four or five-level Likert scale on the question "is the current office-holder honest?" or via feeling thermometer scale, to be asked prior to treatment assignment.

Label the *belief augmented difference in mean voting intentions* the difference in reported votes when researchers condition on both the treatment and villagers' evaluation of the incumbent prior to the treatment. We obtain:

Proposition 5. *Suppose the randomization is at the individual level and the outcome of interest is individual voting level intention.*

The belief augmented conditional difference in mean voting intentions always permits comparability for corruption treatments.

To illustrate our recommended approach, suppose an individual i in village v has either a good opinion ($\Pi_{ivd} = 1$) or a bad opinion ($\Pi_{ivd} = 0$) of the office-holder. The belief augmented conditional differences in mean voting intentions can be recovered by running the following regression:

$$\begin{aligned} Y_{ivd} =& \delta_v + f_0 + f_1 H_{vd} + f_2 \Pi_{ivd} + f_3 T_{ivd} + f_4 T_{ivd} \times H_{vd} + f_5 T_{ivd} \times \Pi_{ivd} + f_6 H_{vd} \times \Pi_{ivd} \\ &+ f_7 T_{ivd} \times H_{vd} \times \Pi_{ivd} + \varepsilon_{ivd} \end{aligned} \tag{15}$$

Equation 15 allows researchers to recover the effect of novel information relative to a pre-existing baseline. For example, f_3 recovers the effect of bad news (i.e., the incumbent is corrupt) for villagers who hold a bad opinion of their office-holder. In turn, $f_3 + f_4 + f_7$ measures the effect of good news for voters who have a high evaluation of the incumbent to begin with. Again there are not a single effect of bad or good news, but as many as categories of evaluation researchers condition on (in our illustration, two, but it could be four, six, etc.).

These estimates can again be used to bound the effect of information treatment. The effect of providing novel information about the incumbent's corruption in any given environment—b_2 by Equation 5—always satisfies $\min\{f_3, f_3 + f_5\} \leq b_2 \leq \max\{f_3, f_3 + f_5\}$, and similarly for other type of news. Hence, studies which focus on individual-level outcomes are useful both for knowledge accumulation and to evaluate the possible benefits of informational campaign.

While it seems that comparability is easily achieved with individual outcomes using beliefs, let us, however, add a word of caution. The belief augmented difference in means permits comparability *only if* the voters' prior evaluations are being recorded on a sufficiently fine-grained scale. If the scale is too broad to account for the heterogeneity in voters' initial information, estimates will again potentially be a function of the circumstances at the time of the intervention, and comparability is compromised (we develop this point fully in the potential outcome framework in Appendix A).

To conclude this section, let us describe some similarities as well as the main differences between the approach we advocate and the specification adopted in the Metaketa initiative. We are in full agreement with Dunning et al. (2020) regarding the importance of measuring voters' beliefs. We part ways regarding how to use this variable. The studies in Metaketa I use voters' evaluations of office-holders to obtain a consistent definition good and bad news. Good news is defined as information more favorable to the incumbent than the voter's initial opinion; bad news consists of less favorable information. Dunning et al. (2020), thus, provide an innovative solution to a conceptual issues: what is good/bad news in practice (i.e., outside a clean theoretical model)? This approach has many advantages, and a few downsides, whose discussion is beyond the scope of the present work. More importantly for us, such operationalization of good and bad news does little to solve the comparability issue we identify in this paper. Each study yields an estimate of a circumstance-specific estimand (each estimate is, again, a weighted average where each weight represents the proportion of voters observing each given level of performance, a function, in turn, of the economic circumstances). In contrast, our approach does not define good or bad news relative to voters' initial evaluation of the incumbent. Rather, it proposes to use voters' priors to identify stable control groups across studies conducted in similar contexts, thus ensuring comparability.

8 Conclusion

"When you can measure what you are speaking about, you know something about it; when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be." This statement by Lord Kevin (1889) highlights the importance of empirical analysis in the production of knowledge. With the recent causal revolution, this is as true as ever. But cumulative knowledge in social sciences does not just require to obtain unbiased and accurate estimate of an underlying quantity. It also necessitates that studies all measure the same estimand. Studies must employ an empirical specification that permits comparability.

Using informational campaigns and their impact on electoral accountability as example, we highlight that comparability should not be assumed. The conditional difference in mean vote shares for village-level randomization, or voting intentions/choices for individual-level randomization, commonly used in the literature, often fails to meet this standard. It is likely to yield an estimate of the effect of providing new information to voters in specific circumstances (such as, for example, the state of the economy at the time when the study takes place). We offer recommendation to recover comparability. If researchers have access to all relevant pieces of information available to voters absent their intervention, a hard task, then, we show, all studies that condition on say information yield estimates of the same underlying estimates. We also highlight a more practical fix: studying individual outcomes and conditioning on voters' opinion of their representative prior to randomization. We also discuss how results from such analysis can inform policy-makers' decision whether to launch an informational campaign.

As a final note, we would like to reiterate the demarcation between the quest for unbiased estimates and the search for comparable estimates. The first, the subject of the causal revolution, requires to obtain perfect balance between treated and control units *within a sample*. It regards the collection, and creation, of observations. The second, the one we studied here, seeks to achieve perfect balance for treated units and for control units, respectively, *across samples*. It concerns the analysis of the data. Both are complementary objectives and, we argue here, essential for the production of knowledge in social sciences.

References

Adida, Claire, Jessica Gottlieb, Eric Kramon, and Gwyneth McClendon. Forthcoming. “Under What Conditions Does Performance Information Influence Voting Behavior? Lessons from Benin.” in Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, and Craig McIntosh (Eds.). *Metaketa I: The Limits of Electoral Accountability.* Cambridge University Press.

Arias, Eric, Horacio A. Larreguy, John Marshall and Pablo Querubin. Forthcoming. “When Does Information Increase Electoral Accountability? Lessons from a Field Experiment In Mexico.” in Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, and Craig McIntosh (Eds.). *Metaketa I: The Limits of Electoral Accountability.* Cambridge University Press.

Ashworth, Scott and Kenneth W. Shotts. 2010. “Does informative media commentary reduce politicians’ incentives to pander?.” Journal of Public Economics 94(11-12): 838-847.

Ashworth, Scott. 2012. “Electoral accountability: recent theoretical and empirical work.” Annual Review of Political Science 15: 183-201.

Ashworth, Scott and Ethan Bueno de Mesquita. 2014. “Is voter competence good for voters?: Information, rationality, and democratic performance.” American Political Science Review. 108(3): 565-587.

Ashworth, Scott, Ethan Bueno de Mesquita, and Amanda Friedenberg. 2017. “Accountability and information in elections.” American Economic Journal: Microeconomics 9(2): 95-138.

Ashworth, Scott, Ethan Bueno de Mesquita, and Amanda Friedenberg. 2018. “Learning about voter rationality.” American Journal of Political Science 62(1): 37-54.

Banerjee, Abhijit V., and Esther Duflo. ”The experimental approach to development economics.” Annual Review of Economics 1.1 (2009): 151-178.

Bhandari, Abhit, Horacio Larreguy, and John Marshall. 2018. “An Empirical Anatomy of Political Accountability: Experimental Evidence from a Pre-Election Information Dissemination Campaign in Senegal.” Working paper.

Boas, Taylor C., F. Daniel Hidalgo and Marcus A. Melo. Forthcoming. “Horizontal but Not Vertical: Accountability Institutions and Electoral Sanctioning in Northeast Brazil.” in Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, and Craig McIntosh (Eds.). *Metaketa I: The Limits of Electoral Accountability.* Cambridge University Press.

Bueno de Mesquita, Ethan and Scott A. Tyson. 2020. “The Commensurability Problem: Conceptual Difficulties in Estimating the Effect of Behavior on Behavior.” American Political Science Review 114(2): 375-391.

Buntaine, Mark, Sarah Bush, Ryan Jablonski, Dian Nielson and Paula Pickering. Forthcoming. “Budgets, SMS Texts, and Votes in Uganda.” in Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, and Craig McIntosh (Eds.). *Metaketa I: The Limits of Electoral Accountability.* Cambridge University Press.

Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, and Craig McIntosh. Forthcoming. "Do Informational Campaigns Promote Electoral Accountability?", in Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, and Craig McIntosh (Eds.). *Metaketa I: The Limits of Electoral Accountability*. Cambridge University Press.

Eggers, Andrew C. "Quality-based explanations of incumbency effects." The Journal of Politics 79(4): 1315-1328.

Ferraz, Claudio and Frederico Finan. 2008. "Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes." Quarterly Journal of Economics 123(2):703–745.

Fox, Justin. 2007. "Government transparency and policymaking." Public choice 131(1-2): 23-44.

Fox, Justin and Richard Van Weelden. 2012. "Costly transparency." Journal of Public Economics 96(1-2): 142-150.

Grossman, Guy, Kristin Michelitch, and Carlo Prato. 2018. " Candidate entry and vote choice in the wake of incumbent performance transparency initiatives." Working Paper.

Hacking, Ian, and Jan Hacking. Representing and intervening: Introductory topics in the philosophy of natural science. Cambridge University Press, 1983.

Larreguy, Horacio, John Marshall, and James M. Snyder. "Publicizing malfeasance: When the local media structure facilitates electoral accountability in Mexico." The Economic Journal (2020).

. Lord Kelvin, Sir William Thomson. 1889. "" in *Popular lectures and address*: 73-136. London, UK: MacMillan and Co.

Lierl, Malte and Marcus Holmlund. Forthcoming. "Performance-Based Voting in Local Elections: Experimental Evidence from Burkina Faso." in Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, and Craig McIntosh (Eds.). *Metaketa I: The Limits of Electoral Accountability*. Cambridge University Press.

Platas, Melina and Pia Raffler. Forthcoming. "Meet the Candidates: Field Experimental Evidence on Learning from Politician Debates in Uganda." in Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, and Craig McIntosh (Eds.). *Metaketa I: The Limits of Electoral Accountability*. Cambridge University Press.

Prat, Andrea. 2005. "The wrong kind of transparency." American Economic Review 95(3): 862-877.

Sircar, Neelanjan and Simon Chauchard. 2020. "Dilemmas and Challenges of Citizen Information Campaigns: Lessons from a Failed Experiment in India" in Dunning, Thad, Guy Grossman, Macartan Humphreys, Susan D. Hyde, and Craig McIntosh (Eds.). *Metaketa I: The Limits of Electoral Accountability*. Cambridge University Press.

Wolton, Stephane. 2019."Are Biased Media Bad for Democracy?." American Journal of Political Science 63(3): 548-562.

Appendix

A Comparability: a potential outcome approach

In this appendix, we show how the comparability issues we uncovered with the help of a formal model can also be understood using the potential outcome framework. We look first at village-level outcome and then at individual outcomes.

Village-level outcome

We first introduce some notation. Let $T \in \{0, 1\}$ denote treatment ($T = 1$) or no treatment ($T = 0$). Let $N \in \{N^+, N^-\}$ be the type of news revealed in the treatment N^+ indicating good news (information which raises voters' posteriors) and N^- indicating bad news (information which decreases the voters' posteriors). In addition to the treatment, villages differ in other dimensions as well. Let $O \in \{0, 1\}$ be a set of observable characteristics which for simplicity takes two values 0 or 1. These are factors that can be measured by the researchers (e.g., public good project outcomes in villages). And let $U \in \{0, 1\}$ be a set of unobservable factors, which again take a binary form. These factors cannot be measured by the researcher (e.g., individual signals about the state of the economy).

We suppose that the distributions of observable and unobservable factors can depend on the type of news provided as well the circumstances at the time of the study, captured by the variable θ. We introduce two useful distributions that play a role below. For a given θ and type of news $n \in \{N^-, N^+\}$, let $\xi_{OU}(\theta, n)$ the joint distribution of observable and unobservable factors in the population (e.g., $\xi_{00}(\theta, n)$ is the proportion of villages that have observable factors set at 0 and unobservable factors set at 0 for a given set of circumstances θ and news n). To take the example of our model, this corresponds to the distribution of successful/failed public good projects as a function of the economic circumstances. Let $\phi_{U|o}(\theta, n)$ the conditional distribution of unobservable factors given a value $o \in \{0, 1\}$ of observable factors. The distributions are independent of specific circumstances θ if and only if $\xi_{OU}(\theta, n) = \xi_{OU}(\theta', n)$, $\phi_{U|O}(\theta, n) = \phi_{U|O}(\theta', n)$ for all $O, U \in \{0, 1\}^2$, all $\theta \neq \theta'$, and all $n \in \{N^+, N^-\}$.

When it comes to the dependent variable, let $Y_v(T, N, O, U)$ be the incumbent's vote share in a village v depending on the treatment (T and N), observable factors (O), and unobservable factors (U). In turn, let Y_v be the observed incumbent's vote share in village v. Like in the main text, we assume perfect random assignment of the treatment and that infinite number of observations are available (so that the mean equals the expectation).

Let $\tau^{cond}(\cdot)$ be the estimand measured using the conditional difference in means vote share, where the condition is on the type $n \in \{N^+, N^-\}$ of news. The main question is whether τ^{cond} is or not a function of the precise circumstances θ in which the study takes place. This estimand is:

$$
\begin{aligned}
& E(Y_v|T=1,N=n) - E(Y_v|T=0,N=n) \\
=& Pr(O=0,U=0|n)\big(E(Y_v|T=1,N=n,O=0,U=0) - E(Y_v|T=0,N=n,O=0,U=0)\big) \\
& + Pr(O=1,U=0|n)\big(E(Y_v|T=1,N=n,O=1,U=0) - E(Y_v|T=0,N=n,O=1,U=0)\big) \\
& + Pr(O=0,U=1|n)\big(E(Y_v|T=1,N=n,O=0,U=1) - E(Y_v|T=0,N=n,O=0,U=1)\big) \\
& + Pr(O=1,U=1|n)\big(E(Y_v|T=1,N=n,O=1,U=1) - E(Y_v|T=0,N=n,O=1,U=1)\big) \\
=& \xi_{11}(\theta,n)\big(E(Y_v(1,n,1,1) - E(Y_v(0,n,1,1)\big) + \xi_{01}(\theta,n)\big(E(Y_v(1,n,0,1) - E(Y_v(0,n,0,1)\big) \\
& + \xi_{10}(\theta,n)\big(E(Y_v(1,n,1,0) - E(Y_v(0,n,1,0)\big) + \xi_{00}(\theta,n)\big(E(Y_v(1,n,0,0) - E(Y_v(0,n,0,0)\big) \\
\equiv& \tau^{cond}(\theta,n)
\end{aligned}
$$

The decomposition above indicates again that, with an infinite number of observations, the treated and control groups are well balanced in term of distributions of observable and unobservable factors. Second, the estimand $\tau^{cond}(\theta, n)$ is again a weighted average of four components. These components are the effect of the treatment for all possible combinations of observable characteristics $o \in \{0, 1\}$ and unobservable factors $u \in \{0, 1\}$: $E(Y_v(T = 1, N = n, O = o, U = u)) - E(Y_v(T = 0, N = n, O = o, U = u))$. These components are all independent from the circumstances. However, as we indicated, the estimand is not a simple average, but a weighted one. And each weight for these different components is possibly a function of the precise realized circumstances θ: $\xi_{ou}(\theta, n)$. Hence, the estimand is circumstance-specific. Unless the joint distribution of observable and unobservable factors is independent of θ, whenever circumstances change (e.g., the survey takes place in good or bad economic conditions), so does the estimand. The conditional difference in mean vote shares, thus, does not necessarily permit comparability.

What about the augmented conditional difference in mean vote shares, where researchers condition on the value of observable outcomes $o \in \{0,1\}$ in addition to the type of news n. The estimand, $\tau^{aug}(\cdot)$, then equals:

$$
\begin{aligned}
&E(Y_v|T=1,N=n,O=o) - E(Y_v|T=0,N=n,O=o)\\
=&Pr(U=0|O=o,n)\big(E(Y_v|T=1,N=n,O=o,U=0) - E(Y_v|T=0,N=n,O=o,U=0)\big)\\
&+Pr(U=1|O=o,n)\big(E(Y_v|T=1,N=n,O=o,U=1) - E(Y_v|T=0,N=n,O=o,U=1)\big)\\
=&\phi_{0|o}(\theta,n)\big(E(Y_v(1,n,o,0) - E(Y_v(0,n,o,0)\big) + \phi_{1|o}(\theta,n)\big(E(Y_v(1,n,o,1) - E(Y_v(0,n,o,1)\big)\\
\equiv&\tau^{aug}(\theta,n,o)
\end{aligned}
$$

Again, the estimand is a function of the environment θ unless the conditional distributions of unobservable factors in the population $\phi_{U|o}(\cdot)$ is independent of θ. One of two conditions need to be met for this independence to hold: first, statistical independence or second, U is fully correlated with O (so $\phi_{U=o|o}(\theta,n)=1$). While the conditions for the augmented conditional difference to permit comparability are milder than for the simple conditional difference in means, these conditions are unlikely to be met in many situations (e.g., voters have some information about the realized underlying environment, say the state of the economy).

Individual outcomes

We keep much of the notation above, but introduce subscript i to designate individuals, with v referring to villages as above. The treatment is now at the individual level $T_i \in \{0,1\}$. We suppose that observable characteristics are village specific $O_v \in \{0,1\}$ (e.g., successful public good project or failure). Unobservable factors are supposed to be individual specifics: $U_i \in \{0,1\}$ (e.g., whether the individual got some additional information about the office-holder or the economic environment). We slightly adapt the notation above (adding subscripts to reflect the distinction village/individual). We denote $\xi_{O_vU_i}(\theta,n)$ the joint distribution of observable and unobservable factors and $\phi_{U_i|O_v}(\theta,n)$ the distribution of unobservable factors $U_i \in \{0,1\}$ among villagers conditional on observable factor

$O_v \in \{0,1\}$ at the village level as a function of the circumstances and, possibly, the type of news provided (e.g., individuals could get an unobserved signal of the incumbent's type so U_i would depend on good/bad news for corruption treatments). Let $\psi_{O_v}(\theta, n)$ be the marginal distribution of O_v across villages as a function of θ and n (e.g., the distribution of success/failure as a function of the economic circumstances and the incumbent's type as in our model).

We adapt the dependent variable to individual level outcome. Let $Y_i(T_i, N, O_v, U_i)$ be the an individual's vote choice (or intention) as a function of (T and N), observable factors (O_v) in her village, and individual unobservable factors (U_i). In turn, let Y_i be the observed or reported individual's vote choice/intention. Again, we assume perfect random assignment of the treatment and that infinite number of observations are available for individuals within villages (treated/control) and for villages in the sample (so that the mean equals the expectation).

In a village with outcome $O_v = o_v \in \{0,1\}$, the impact of the treatment is

$$\begin{aligned}
&E(Y_i|T_i = 1, N = n, O_v = o_v) - E(Y_i|T = 0, N = n, O_v = o_v)\\
=&Pr(U_i = 0|n, O_v = o_v)\big(E(Y_i|T = 1, N = n, O_v = o_v, U_i = 0) - E(Y_i|T = 0, N = n, O_v = o_v, U_i = 0)\big)\\
&+ Pr(U_i = 1|n, O_v = o_v)\big(E(Y_i|T = 1, N = n, O_v = o_v, U_i = 1) - E(Y_v|T = 0, N = n, O_v = o_v, U_i = 1)\big)\\
=&\phi_{0|o_v}(\theta, n)\big(E(Y_i(1, n, O_v = o_v, 0) - E(Y_i(0, n, O_v = o_v, 0)\big)\\
&+ \phi_{1|o_v}(\theta, n)\big(E(Y_i(1, n, O_v = o_v, 1) - E(Y_i(0, n, O_v = o_v, 1)\big)\\
&\equiv \nu(\theta, n, o_v)
\end{aligned}$$

Across villages, the conditional difference in means estimates the following estimand, denoted $\tau_{ind}^{cond}(\cdot)$:

$$\begin{aligned}
&E(Y_i|T_i = 1, N = n) - E(Y_i|T = 0, N = n)\\
&= \psi_0(\theta, n)\nu(\theta, n, 0) + \psi_1(\theta, n)\nu(\theta, n, 1)\\
&= \sum_{o_v \in \{0,1\}} \psi_{o_v}(\theta, n) \begin{pmatrix} \phi_{0|o_v}(\theta, n)\big(E(Y_i(1, n, O_v = o_v, 0) - E(Y_i(0, n, O_v = o_v, 0)\big) \\ +\phi_{1|o_v}(\theta, n)\big(E(Y_i(1, n, O_v = o_v, 1) - E(Y_i(0, n, O_v = o_v, 1)\big) \end{pmatrix}\\
=&\xi_{11}(\theta, n)\big(E(Y_v(1, n, 1, 1) - E(Y_v(0, n, 1, 1)\big) + \xi_{01}(\theta, n)\big(E(Y_v(1, n, 0, 1) - E(Y_v(0, n, 0, 1)\big)\\
&+ \xi_{10}(\theta, n)\big(E(Y_v(1, n, 1, 0) - E(Y_v(0, n, 1, 0)\big) + \xi_{00}(\theta, n)\big(E(Y_v(1, n, 0, 0) - E(Y_v(0, n, 0, 0)\big)\\
\equiv&\tau_{ind}^{cond}(\theta, n)
\end{aligned}$$

As noted in the text, there is nothing special about individual outcomes. The estimand measured using the individual-level conditional difference in mean voting intentions looks very much like the estimand estimated using the village-level conditional difference in mean vote shares. Thus, it is plagued by the same problem. Comparability fails to be achieved unless the distributions of observed and unobserved factors are independent of the precise circumstances in which the study takes place.

Let us turn to the belief-augmented conditional difference in mean voting intentions. In our cases, individuals have four different beliefs, one associated with an observable and an unobservable factor. Let's then denote individual i's belief about the office-holder's honesty π_i. We suppose that $\pi_i = \pi(O_v, U_i)$; that is, her belief is a function of observable and unobservable factors, and solely of these. Further, we assume that all individuals have the same belief conditional on observing the same information. For simplicity, but without loss of generality, we assume that $\pi(1, 1) > \pi(1, 0) > \pi(0, 1) > \pi(0, 0)$. The researcher uses a four-level Lickert scale L. The answer of individual i is denoted $l_i \in \{1, 2, 3, 4\}$ with 1 corresponding to $\pi_i(0, 0)$, 2 to $\pi_i(0, 1)$, etc. Now in a given village, the impact of the treatment conditioning on a particular $l \in \{1, 2, 3, 4\}$ is:

$$\begin{aligned}
&E(Y_i|T_i = 1, N = n, O_v = o_v, l_i = l) - E(Y_i|T = 0, N = n, O_v = o_v, l_i = l) \\
=&E(Y_i|T_i = 1, N = n, O_v = o_v, \pi_i = \pi(o_v, u_i)) - E(Y_i|T = 0, N = n, O_v = o_v, \pi_i = \pi(o_v, u_i)) \\
=&E(Y_i|T_i = 1, N = n, O_v = o_v, U_i = u_i) - E(Y_i|T = 0, N = n, O_v = o_v, U_i = u_i) \\
=&E(Y_i(1, n, O_v = o_v, U_i = u_i) - E(Y_i(0, n, O_v = o_v, U_i = u_i)\big) \\
&\equiv \nu^{Bcond}(o_v, u_i)
\end{aligned}$$

Across villages, the belief augmented conditional difference in means estimates the following estimand, denoted $\tau_{ind}^{Bcond}(\cdot)$, is simply

$$\nu^{Bcond}(o_v, u_i) = \tau_{ind}^{Bcond}(o_v, u_i)$$

There are, obviously, four estimands, one for each pair of $o_v \in \{0, 1\}$ and $u_i \in \{0, 1\}$, and all do not depend on the realized circumstances θ. Hence, the belief augmented conditional difference in mean voting intentions permits comparability.

As a final note, let us stress the importance of using the right scale to condition on belief. Suppose that researchers use a two-level Lickert scale with low approval ($l = 1$) or high approval ($l = 2$) of the incumbent. Assume, for ease of exposition, that individuals with beliefs $pi(O_v = 1, U_i = 1)$ and $\pi(1, 0)$ respond $l_i = 2$ and those with other beliefs ($\pi(0, 1)$ and $\pi(0, 0)$) respond $l_i = 1$. It is immediate that conditioning on a particular response is equivalent to conditioning on a specific value of observable outcomes, such as project success/failure (other splits obviously give other types of condition). Hence, in this case, the belief augmented conditional difference in mean voting intentions is the same as the augmented conditional difference in means. And it, thus, suffers from the same problem. It permits comparability only if unobserved and observed factors are fully correlated or unobserved factors do not depend on circumstances. This is why fine grained measures of individuals' evaluation of their representative (prior to treatment) should be preferred to broad categories.

B Proofs

Proof of Proposition 1

Recall that, for the performance treatment, the effect of bad news (a_2) and good news ($a_2 + a_3$) are, respectively:

$$a_2 =(1-\lambda)\Big(G\big((\pi^C-\pi^I)(\alpha(h)-\alpha(c))\gamma^e\big) - G\big((\pi^C-\mu^v(0,0))(\alpha(h)-\alpha(c))\gamma^e\big)\Big) \tag{B.1}$$

$$\text{and} \quad a_2 + a_3 =(1-\lambda)\Big(G\big((\pi^C-\pi^I)(\alpha(h)-\alpha(c))\gamma^e\big) - G\big((\pi^C-\mu^v(1,1))(\alpha(h)-\alpha(c))\gamma^e\big)\Big) \tag{B.2}$$

The effect of the performance treatment is zero whenever $\lambda = 1$. Suppose then $\lambda < 1$. Given $\alpha(h) > \alpha(c)$, voters' posteriors satisfies: $\mu^(0,0) < \pi^I < \mu^v(1,1)$. Hence, $G\big((\pi^C-\pi^I)(\alpha(h)-\alpha(c))\gamma^e\big) < G\big((\pi^C-\mu^v(0,0))(\alpha(h)-\alpha(c))\gamma^e\big)$ and $G\big((\pi^C-\pi^I)(\alpha(h)-\alpha(c))\gamma^e\big) > G\big((\pi^C-\mu^v(1,1))(\alpha(h)-\alpha(c))\gamma^e\big)$. Therefore, $a_2 < 0$ and $a_2 + a_3 > 0$.

Now, for the corruption treatment the effect of good news and bad news are, respectively

$$\begin{aligned} b_2 =&\lambda\alpha(h)\gamma(\theta)\Big(G\big((\pi^C-\mu_v(1,\emptyset))(\alpha(h)-\alpha(c))\gamma^e\big) - G\big(\pi^C(\alpha(h)-\alpha(c))\gamma^e\big)\Big) \\ &+\lambda\big(1-\alpha(c)\gamma(\theta)\big)\Big(G\big((\pi^C-\mu_v(0,\emptyset))(\alpha(h)-\alpha(c))\gamma^e\big) - G\big(\pi^C(\alpha(h)-\alpha(c))\gamma^e\big)\Big) \\ &+(1-\lambda)\Big(G\big((\pi^C-\pi^I)(\alpha(h)-\alpha(c))\gamma^e\big) - G\big(\pi^C(\alpha(h)-\alpha(c))\gamma^e\big)\Big) \end{aligned} \tag{B.3}$$

$$\begin{aligned} \text{and} \quad b_2 + b_3 =&\lambda\alpha(c)\gamma(\theta_c,\theta_s,\theta_d)\Big(G\big((\pi^C-\mu_v(1,\emptyset))(\alpha(h)-\alpha(c))\gamma^e\big) - G\big((\pi^C-1)(\alpha(h)-\alpha(c))\gamma^e\big)\Big) \\ &+\lambda\big(1-\alpha(c)\gamma(\theta_c,\theta_s,\theta_d)\big)\Big(G\big((\pi^C-\mu_v(0,\emptyset))(\alpha(h)-\alpha(c))\gamma^e\big) - G\big((\pi^C-1)(\alpha(h)-\alpha(c))\gamma^e\big)\Big) \\ &+(1-\lambda)\Big(G\big((\pi^C-\pi^I)(\alpha(h)-\alpha(c))\gamma^e\big) - G\big((\pi^C-1)(\alpha(h)-\alpha(c))\gamma^e\big)\Big) \end{aligned} \tag{B.4}$$

Given $\mu^v(\cdot,\tau_{I_d}=h)=1$ and $\mu^v(\cdot,\tau_{I_d}=c)=0$, we necessarily have $\mu^v(\cdot,\tau_{I_d}=h) > \pi^I > \mu^v(\cdot,\tau_{I_d}=c)$ as well as $\mu^v(\cdot,\tau_{I_d}=h) \geq \mu^v(\omega_v^1=1,\emptyset) > \mu^v(\omega_v^1=0,\emptyset) \geq \mu^v(\cdot,\tau_{I_d}=c)=0$, with the first inequality strict whenever $\alpha(c) > 0$ and the second strict whenever $\alpha(h) < 1$. If $\lambda < 1$, then necessarily $b_2 < 0 < b_2 + b_3$. If $\lambda = 1$, then $b_2 \leq 0$ with strict inequality whenever $\alpha(h) < 1$ and $b_2 + b_3 \geq 0$ with strict inequality whenever $\alpha(c) > 0$. $\square$

Proof of Proposition 2

Simple observation of Equation B.1 and Equation B.2 yields that the (limit) estimates from the performance treatment only depends on invariant primitives of the model (π^I, π^C, γ^e, $\alpha(c)$, $\alpha(h)$, λ, and $G(\cdot)$). Hence, the estimates recovered from samples in different places or at different points in times all measure the same estimand.

For the corruption treatment, using Equation B.3 and Equation B.4, the estimate will measure a different estimand from one random sample to the next as the realizations of the environment change (the exact values of the θ's change). Hence, the estimand being estimated across different samples is the same only if the economic environment does not influence the incumbent's performance: $\gamma(\underline{\theta}) = \gamma(\overline{\theta})$.

Proof of Proposition 3

The proof follows from a simple inspection of Equation 11 and Equation 12, which shows that the (limit) estimates are only a function of the primitives, and not on the economic shock realization. Thus, fixing the context studies yield estimates of the same estimand. $\square$

Proof of Proposition 4

Turning to individual=level outcomes, we need a few additional pieces of notation to study the theoretical equivalent of the conditional difference in means. Suppose that within a village, a proportion $\eta \in (0, 1)$ of villagers are treated. We now denote $\iota_i \in \{\emptyset, \omega_v^1, \tau_{I_d}\}$ the information provided to villager i by researchers (with as before $\emptyset$ meaning no treatment, ω_v^1 the performance treatment, and τ_{I_d} the corruption treatment). Using Equation 2, we can determine the probability that a randomly drawn villager i votes in favour of the incumbent. Denoting it $S_v^i(z_i, \iota_i)$ (with

$z_i \in \{\emptyset, \omega_v^1\}$ the information provided by Nature), it is equal to

$$S_v^i(z_i, \iota_i) = 1 - G\big((\pi^C - \mu_v(z_i, \iota_i)(\alpha(h) - \alpha(c))\gamma^e\big) \tag{B.5}$$

The posteriors for a voter are defined in Table 1 above. We can then use the same analysis as in Section 4 to determine the average voting intention among treated (taking advantage of the mass of villagers in each village). These averages are for the performance and corruption treatments, respectively:

$$E^i\big(S_v^i(z_i, \iota_i)|\iota_i = \omega_v^1\big) = 1 - G\big((\pi^C - \mu_v(\emptyset, \omega_v^1)(\alpha(h) - \alpha(c))\gamma^e\big) \tag{B.6}$$

$$E\big(S_v^i(z, \iota_i)|\iota_i = \tau_{I_d}\big) = 1 - G\big((\pi^C - \mu_v(\emptyset, \tau_{I_d})(\alpha(h) - \alpha(c))\gamma^e\big) \tag{B.7}$$

In turn, among the villagers belonging to the control group, recall that a proportion λ learns the project outcome in their village via Nature. Given that we have a mass of villagers, the average reported vote for the incumbent is:

$$E^i\big(S_v^i(z, \iota_i)|\iota_i = \emptyset\big) = 1 - \lambda G\big((\pi^C - \mu_v(\omega_v^1, \emptyset))(\alpha(h) - \alpha(c))\gamma^e\big) - (1-\lambda)G\big((\pi^C - \pi^I)(\alpha(h) - \alpha(c))\gamma^e\big) \tag{B.8}$$

Proceeding as we did in Section 5, within a village, the effects of the performance treatment estimated in Equation 13—d_2 for villagers in a village where the project fails ($\Omega_{vds} = 0$), $d_2 + d_3$ for villagers in a village where the project succeeds ($\Omega_{vds} = 1$)—correspond to:

$$d_2 = (1 - \lambda)\Big(G\big((\pi^C - \pi^I)(\alpha(h) - \alpha(c))\gamma^e\big) - G\big((\pi^C - \mu_v(0,0))(\alpha(h) - \alpha(c))\gamma^e\big)\Big) \tag{B.9}$$

$$d_2 + d_3 = (1 - \lambda)\Big(G\big((\pi^C - \pi^I)(\alpha(h) - \alpha(c))\gamma^e\big) - G\big((\pi^C - \mu_v(1,1))(\alpha(h) - \alpha(c))\gamma^e\big)\Big) \tag{B.10}$$

Unsurprisingly (given our assumptions that researchers draw a mass of voters), individual-level effects are exactly equivalent to the estimates obtained in village-level analyses. Performance treatments thus always permit comparability, even when the analysis is at the individual level.

Similarly, we can verify that for corruption treatment the effect estimated in Equation 14 corresponds to

$$
\begin{aligned}
e_2 =&(1-\lambda)\Big(G\big((\pi^C-\pi^I)(\alpha(h)-\alpha(c))\gamma^e\big)-G\big(\pi^C(\alpha(h)-\alpha(c))\gamma^e\big)\Big)\\
&+\lambda\alpha(c)\gamma(\theta)\Big(G\big((\pi^C-\mu_v(1,\emptyset))(\alpha(h)-\alpha(c))\gamma^e\big)-G\big(\pi^C(\alpha(h)-\alpha(c))\gamma^e\big)\Big)\\
&+\lambda\big(1-\alpha(c)\gamma(\theta)\big)\Big(G\big((\pi^C-\mu_v(0,\emptyset))(\alpha(h)-\alpha(c))\gamma^e\big)-G\big(\pi^C(\alpha(h)-\alpha(c))\gamma^e\big)\Big)
\end{aligned}
\tag{B.11}
$$

$$
\begin{aligned}
\text{and}\quad e_2+e_3 =&(1-\lambda)\Big(G\big((\pi^C-\pi^I)(\alpha(h)-\alpha(c))\gamma^e\big)-G\big((\pi^C-1)(\alpha(h)-\alpha(c))\gamma^e\big)\Big)\\
&+\lambda\alpha(h)\gamma(\theta)\Big(G\big((\pi^C-\mu_v(1,\emptyset))(\alpha(h)-\alpha(c))\gamma^e\big)-G\big((\pi^C-1)(\alpha(h)-\alpha(c))\gamma^e\big)\Big)\\
&+\lambda\big(1-\alpha(h)\gamma(\theta)\big)\Big(G\big((\pi^C-\mu_v(0,\emptyset))(\alpha(h)-\alpha(c))\gamma^e\big)-G\big((\pi^C-1)(\alpha(h)-\alpha(c))\gamma^e\big)\Big)
\end{aligned}
\tag{B.12}
$$

As for the case of aggregate-level outcomes, corruption treatments do not permit comparability whenever $\gamma(\underline{\theta}) \neq \gamma(\overline{\theta})$. □

Proof of Proposition 5

In Equation 15, Π denotes voter's evaluations of the incumbent prior to the treatment being administered. Within our framework this is a function of the distribution of politicians' types (i.e., the voters' prior) and, potentially, the incumbent's performance (project outcome in the village). The distribution of types is constant at the country level, so conditioning on voter's evaluation is equivalent to conditioning on the incumbent's performance. Thus, the estimates in Equation 15 correspond to those obtained under the augmented conditional difference in means for village-level outcomes (as shown in the previous preposition, there is a full equivalence between individual and aggregate-level analysis in our setting). As such, as stated in Proposition 3, the estimates obtained by adopting this specification are fully comparable.