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

Federica Izzo Torun Dewan Stephane Wolton

Current draft: March 18, 2022

Link to the most recent version

Abstract

Two (non-exhaustive) conditions are necessary for knowledge accumulation: unbiasedness and comparability. Research designs should be unbiased so that researchers obtain correct estimates of an underlying quantity. Empirical specifications 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 highlight the difficulty to obtain comparability even after removing all concerns linked to external validity, all statistical noise, and all sources of bias. Commonly used specifications reach comparability only under specific, non-testable conditions. We propose several recommendations to restore comparability.

Word count: approx. 9,650 words (Monterrey word count)

Key words: electoral accountability, accumulation of knowledge, comparability, bias, theoretical implications of empirical models

*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

How do we know? In social sciences, two broad perspectives dominate. The first view emphasizes the role of theory. We achieve knowledge accumulation, i.e., we uncover stable phenomena, 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 perspective 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 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. Knowledge accumulation also requires comparability, i.e., the basic idea that the empirical specification employed yields an estimate of the same estimand for all possible draws of observations. Knowledge accumulation can be achieved only if empirical studies measure the same quantity.

Our notion of comparability is broader than the commonly recognized problem of external validity. To highlight this, our paper makes an important 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. Thus, we 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).

External validity of an intervention fails when the underlying quantities scholars measure change with the context in which the studies take place. If an intervention is not successful across different settings, it is accepted that it should not be recommended to policy-makers. The estimands

researchers estimate, however, can also be a function of the circumstances. The latter is not a mere problem of heterogeneous effect. 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. When studies measure circumstance-specific estimands, the search for stable phenomena is severely impaired. Knowledge accumulation becomes hopeless. Discussion of external validity—how well findings travel from one context to the next—only makes sense when the underlying quantities of interests are free of influence from always-varying circumstances.

In this paper, we argue that commonly used empirical approaches may fail to measure the same estimand even though the context remains invariant between two interventions. Differences in results may have wrongly been attributed to lack of external validity, when they could have resulted from lack of comparability resulting from the empirical specifications employed. We also detail several recommendations to ensure comparability within, at least, the same setting.

We illustrate our argument with the example of informational campaigns and their consequences for electoral outcomes. There are several reasons for this choice. First, the topic has generated considerable interest theoretically (see, among many others, Ashworth, 2012; Ashworth et al., 2017) and empirically (see, for a recent review, Incerti, 2020). It is, for example, the theme of the first set of coordinated studies under the Metaketa Initiative (Dunning et al., 2020). Second, despite the many studies on the effect of information, no definitive conclusion has yet been reached. As Bhandari, Larreguy, and Marshall (2021: 2) explain, "recent studies identifying the effects of informational campaigns on electoral accountability (...) yield mixed findings."

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 sample a representative mass of observations before randomly dividing the sample between treated and control units. We then ask whether different studies operating in the same context (all studies draw samples from the same setting, with the same underlying fundamentals) and employing the same specification are comparable. In term of specification, we follow the literature and consider the conditional difference in mean vote shares. That is, we assume that researchers compare the average in the treatment group for a certain treatment value (e.g., the incumbent has performed well) with the average among units who would have received the same information, should they have been treated. By design, the only difference between studies is that they occur in different

circumstances (e.g., under specific time-varying economic conditions). We show that the conditional difference in means does not permit comparability.

To see why, consider the following thought experiment, which constitutes a special case of our general 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 so that interventions take place under different economic circumstances. Suppose that 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 this affect the evaluation of the informational campaign assuming voters prefer honest politicians to corrupt ones? 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. In the first survey, the office-holder's performance is high and voters have a relative good baseline opinion of their representative. In the second study, incumbents' performance is low and, absent treatment, constituents evaluate poorly their representative. Think now about the effect of providing information that indicate or reveal that the incumbent is honest. In the first study, this moves voters' posteriors slightly upward from a relatively high baseline. Comparing treated and control units, researchers then uncover a low effect of the informational campaign. In the second study, good news move voters' evaluation upward from a low baseline. Thus, 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.

More generally, we show that the conditional difference in means vote share permit comparability if and only if, after conditioning on the treatment value, the incumbent's electoral standing is unaffected by the circumstances in which the intervention takes place. This condition proves difficult to test within a single study. Rather than hoping it holds, we provide ways for researchers to recover comparability. To do so requires a change of counterfactuals. Instead of only conditioning on what control units would have observed if treated, researchers should also condition on what the treated

villages would have learned absent treatment. This strategy is relatively easy to implement when individual voters are the unit of analysis. Here, researchers just need to collect constituents' relevant opinions of the incumbent prior to the (possible) provision of information and condition their analysis on voters' evaluations. Indeed, a well known property of constituents' belief about their representative is that it encompasses all the information available to voters prior to the intervention. Everything related to circumstances is then accounted for. Conditioning on beliefs resolves the exact problem our work identifies.

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. Much progress has been made in this direction with the use of coordinated experiments. This paper, we hope, helps with this endeavor. It tries to advance 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 two formal model, one illustrative example and one general framework. 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.

1 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, such as Eggers (2017) on the properties of regression discontinuity designs, Ashworth et al. (2018) on the electoral effect of random events, Bueno de Mesquita and Tyson (2019) on the commensurability problem.

In this literature, two papers are closest to our work: Slough and Tyson (2021) and Wolton (2019). Slough and Tyson (2021) uncover the underlying assumptions behind meta-analysis, and highlight the conditions under which studies in different contexts measure the same estimand. Further, Slough and Tyson (2021) take a model-free approach and a macro-perspective. In contrast,

we focus on a specific problem, information campaigns, with a micro-approach. The two papers should, thus, be seen as complementary. Wolton (2019) details how estimates of the electoral consequences of biased media are a function of the political environment (e.g., the partisan identity of office-holder) and the media environment (e.g., the partisan identity of media reporting). As such, Wolton (2019) is interested in the issue of external validity (or lack thereof), not comparability per se as in the present work. Unlike both papers, 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 (2022) studies a political agency model with features analogous to our illustrative example below (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 (2022) 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.

2 Failure of comparability: an illustrative example.

In this section, we present a simplified version of our general model (introduced in Section 3 below) to illustrate our argument. We consider a country divided into a mass D of districts, each constituted

of a mass V of villages, each inhabited by a mass 1 of voters. 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.

Each politician J is one of two types: honest, $\tau_J = 1$, or corrupt, $\tau_J = 0$. Villagers do not know the politicians' types at the beginning of the game. However, it is common knowledge that the incumbent is honest with probability q_I and the challenger with probability q_C (identical in all districts for simplicity). In addition, some voters in each village observe a noisy signal of the incumbent's type, s_i^v. In this set-up, we interpret s_v^i as a voter learning whether the incumbent successfully completed a public project in her village, $\omega^v = 1$, or not, $\omega^v = 0$. Thus, we assume that a proportion λ of voters observe their village-specific project outcome: $s_i^v = \omega^v$ (i.e., each voter has a probability λ of learning ω^v). The remaining voters in a village learn nothing directly, in which case we denote $s_i^v = -1$.

To make project outcomes informative about the incumbent's type, we assume that the probability that a project is successful takes the form $Pr(\omega^v = 1) = \alpha(\tau_{I_d}) \times \theta$, with $0 < \alpha(0) < \alpha(1) < 1$, so honest types are more likely to complete the project. In turn, θ is a shock capturing the environment in which the incumbent operates, with higher θ designating a more favorable economic environment (e.g., higher budget). We assume that θ is common to all office-holders in the country (some correlation across districts would be enough for our results to hold). Villagers do not observe the realization of θ, but it is common knowledge that this shock is distributed according to the continuous and strictly increasing cumulative distribution function F over the interval $[\underline{\theta}, \overline{\theta}]$, with $0 < \underline{\theta} < \overline{\theta} < 1$.

Everything else being equal, voters prefer to elect honest politicians. In addition, their evaluation of the two candidates also depends on an idiosyncratic shock, denoted σ_i^J, with $J \in \{I, C\}$, where $\sigma_i^I - \sigma_i^C$ 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). Overall, the voter's payoff from electing politician J is

$$U_J^v = \tau_J + \sigma_i^J. \tag{1}$$

Informational campaigns. We use this framework to model informational campaigns. We think about researchers randomly selecting (a mass of) villages in each district, and then randomly assigning a subset to treatment and the rest to control. Researchers' intervention should be understood broadly. The researchers could be directly responsible for the randomization or they could exploit features of public policies (e.g., the use of random public audits as in Ferraz and Finan, 2004 and 2008). We consider two kinds of interventions, which differ in the nature of the treatment. In the first type of intervention, a proportion ρ of 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, a proportion ρ of 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*.

Our outcome of interest will be the incumbent's vote share in a village. We follow the best practice in the literature, and analyze the effect of information treatments conditional on the information provided (see Dunning et al., 2020; Ferraz and Finan, 2004; Larreguy et al., 2020, among others). We label this empirical approach the *conditional difference in mean vote shares*. In our model, researchers recover two estimates for each treatment: after project completion or project failure for performance treatment and after revelation of corrupt or honest type for corruption treatment. The conditional difference in mean can be performed separately for each treatment value or with a fully interacted model, the approach we will adopt below.

Timing. The game proceeds as follows. 1. In period 1, Nature draws the incumbents' and the challengers' types in all districts. 2. Nature draws θ, determines project outcomes in all villages ω^v, as well as the λ voters who observe ω^v in each village. 3. Researchers randomly select some villages and, among the treated samples, ρ voters receive the informational treatment. 4. The voters observe their idiosyncratic shock σ_i^J $(J \in \{C_d, I_d\})$ and cast their ballot. 5. The politician who receives the most votes in a district is elected. 6. The game moves to period 2 and ends with payoffs being realized. Since only voters are strategic actors, no equilibrium concept is required, besides the usual assumption that voters are (expected) utility maximizers.

Comparability. Building on our discussion in the introduction, we define comparability in the following way.

Definition 1. *An empirical specification permits comparability if it yields an estimate of the same estimand for all possible draws of observations.*

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.

As we noted, comparability may fail due to a lack of external validity. In our framework, this would correspond to researchers intervening in different contexts, with each study uncovering a context-specific effect. The context corresponds to the primitives of the model. The political context is the distribution of types among politicians (q_I and q_C), and the social context consists of the distribution of economic shocks ($F(\cdot)$ over $[\underline{\theta}, \overline{\theta}]$) and the form production function of completed projects ($\alpha(\widetilde{\tau})\widetilde{\theta}$ for all $\widetilde{\tau}$ and $\widetilde{\theta}$). Throughout our analysis, we assume away these concerns by imposing that researchers always intervene in the same context.

This leaves circumstances as the unique impediment to comparability. In our framework, the circumstances correspond to the actual realization of all random variables. The political circumstances correspond to the actual type of the office-holder $\tau \in \{0, 1\}$. The economic circumstances are the actual value of the shock θ. We suppose that researchers intervene at two different points in time, so we allow the circumstances to vary even though we keep the context fixed. If studies measure circumstance-specific estimands, comparability will fail even though external validity is not a concern.

Theoretical implications of empirical models.

In this subsection, we ask whether the conditional difference in mean vote shares permits comparability. To do so, we first describe how this specification can be operationalised in our setting. First, let us introduce a binary variable denoted T_{vd}, which takes a value one if a village v in district d is treated and zero otherwise. Regarding the content of the treatment, whether we consider the performance or the corruption treatment, the treatment value is either one (if the project is completed in the village, if the incumbent is honest) or zero (if the project fails, if the incumbent is corrupt).

Hence, we denote $Z_{vd} \in \{0, 1\}$ the value of the treatment. To measure the conditional difference in mean vote shares at the village level, the researchers then run the following fully interacted model:

$$Y_{vd} = a_0 + a_1 T_{vd} + a_2 Z_{vd} + a_3 T_{vd} \times Z_{vd} + \epsilon_{vd}, \tag{2}$$

with Y_{vd} the vote share of the incumbent in village v in district d and ϵ_{vd} the noise.

With this regression, the researchers recover estimates for two different estimands. Parameter a_1 corresponds to the effect of the treatment in villages treated with treatment value 0 (i.e., informing voters that the project has not been completed for the performance treatment or that the incumbent is corrupt for the corruption treatment). In turn, the sum $a_1 + a_2$ corresponds to the effect of the treatment in villages treated with treatment value 1 (i.e., informing voters that the project has been completed for the performance treatment or that the incumbent is honest for the corruption treatment).

We now use our model to map these empirical quantities into their theoretical equivalent. Recall that, in our framework, researchers randomly draw a mass of villages, so any theoretical outcome is equal to its associated estimand. We can then simply solve for equilibrium behavior and compute vote shares in treated and control villages to verify whether studies are comparable using our definition above.

To this aim, denote $\mu_v(\tau_{I_d}|s_i^v, z^v)$ voter i's posterior belief that the incumbent is a type $\tau_{I_d} \in \{0, 1\}$, as a function of the information she obtains from nature, $s_i^v \in \{\omega^v, -1\}$, and the treatment she may receive from the researchers, z^v, which takes values $z^v \in \{\omega^v, \emptyset\}$ for the performance treatment and values $z^v \in \{\tau_{I_d}, \emptyset\}$ for the corruption treatment (with $z^v = \emptyset$ if a voter is not reached by the researchers). Notice that, under the performance treatment, $\mu_v(\tau_{I_d}|\omega^v, \omega^v) = \mu_v(\tau_{I_d}| - 1, \omega^v) = \mu_v(\tau_{I_d}|\omega^v, \emptyset)$. By Bayes rule, $\mu_v(1|1, \emptyset) = \frac{q_I \alpha(1)}{q_I \alpha(1) + (1 - q_I)\alpha(0)}$ and $\mu_v(1|0, \emptyset) = \frac{q_I(1-\alpha(1)E(\theta))}{q_I(1-\alpha(1)E(\theta)) + (1-q_I)(1-\alpha(0)E(\theta))}$, with $E(\theta)$ the expected value of the environment θ (which is unobserved by voters). For the corruption treatment, we have $\mu_v(\tau_{I_d}|s_i^v, \tau_{I_d}) = 1$ for all possible voters' signals. Voters who learn nothing have to rely on their prior: $\mu_v(1| - 1, \emptyset) = q_I$.

As described above, the voter prefers to elect a competent candidate. However, her evaluation of the incumbent and challenger is also influenced by the idiosyncratic shocks σ_i^I and σ_i^C. Thus,

from equation (1), voter i casts a ballot for the incumbent if and only if

$$\sigma_i^I - \sigma_i^C > -(\mu_v(1|s_i^v, z^v) - q_C) \tag{3}$$

From this, exploiting the assumption that each village is inhabited by a mass of citizens, we can easily compute the incumbent's vote share in each village and treatment group. In a control village, a proportion λ is informed of the project outcome $\omega^v \in \{0, 1\}$, whereas the rest learn nothing and rely on their prior. The incumbent's realized vote share in a village with project outcome $\omega^v \in \{0, 1\}$ is then given by:

$$(1-\lambda)\Big(1 - G(q_C - q_I)\Big) + \lambda\Big(1 - G(q_C - \mu_v(1|\omega^v, \emptyset))\Big) \tag{4}$$

The control group, as a whole, consists of many villages, some represented by honest types ($\tau_{I_d} = 1$), other by corrupt office-holders ($\tau_{I_d} = 0$), some where the project outcome is successful ($\omega^v = 1$), some where it is not ($\omega^v = 0$). Under our assumption of a mass of villages and districts, the proportion of villages for each event is equal to the probability of each event occurring so that, for example, q^I villages have a honest incumbent, and of those $\alpha(1)\theta$ see the project completed, $1 - \alpha(1)\theta$ experience a project failure. Hence, the incumbent's average vote share in the control group is:

$$\begin{aligned}\mathcal{S}(\theta, \emptyset) =& q_I \begin{pmatrix} \alpha(1)\theta\Big((1-\lambda)\big(1 - G(q_C - q_I)\big) + \lambda\big(1 - G(q_C - \mu_v(1|1, \emptyset))\Big) \\ (1-\alpha(1)\theta)\Big((1-\lambda)\big(1 - G(q_C - q_I)\big) + \lambda\big(1 - G(q_C - \mu_v(1|1, \emptyset))\Big) \end{pmatrix} \\ &+ (1-q_I) \begin{pmatrix} \alpha(0)\theta\Big((1-\lambda)\big(1 - G(q_C - q_I)\big) + \lambda\big(1 - G(q_C - \mu_v(1|1, \emptyset))\Big) \\ (1-\alpha(0)\theta)\Big((1-\lambda)\big(1 - G(q_C - q_I)\big) + \lambda\big(1 - G(q_C - \mu_v(1|1, \emptyset))\Big) \end{pmatrix}\end{aligned} \tag{5}$$

Next, we turn to the performance treatment. In this case, treated voters' signals s_i^v are inconsequential: all treated voters have the same posterior (since voters receive the same information from researchers as from Nature). As a result, in a treated village under the performance treatment with value $\omega^v \in \{0, 1\}$, the incumbent's vote share is:

$$(1-\lambda)(1-\rho)\Big(1 - G(q_C - q_I)\Big) + \Big(1 - (1-\lambda)(1-\rho)\Big)\Big(1 - G(q_C - \mu_v(1|-1, \omega^v))\Big). \tag{6}$$

When it comes to the incumbent vote share in the treated group with treatment value $\omega^v \in \{0,1\}$, we again average across a mass of villages. In this case, due to the nature of the treatment, only one type of villages is present in the sample: those that experience project outcome ω^v. Hence, the average vote share for treated villages for value ω^v of the performance treatment is the same as the incumbent's vote share in a single treated village:

$$\mathcal{S}(\theta,\omega^v) = (1-\lambda)(1-\rho)\Big(1-G(q_C-q_I)\Big) + \Big(1-(1-\lambda)(1-\rho)\Big)\Big(1-G(q_C-\mu_v(1|-1,\omega^v))\Big) \quad (7)$$

Finally, for the corruption treatment, the vote share of the incumbent in a treated village with treatment value τ_{I_d} is an average across four groups: the voters who only learn the treatment, the voters who learn either the treatment or the project outcome ω^v in their village, and the voters who learn both. The vote share assumes the following form.

$$\lambda\begin{pmatrix} \rho\Big(1-G(q_C-\mu_v(1|\omega^v,\tau_{I_d}))\Big) \\ +(1-\rho)\Big(1-G(q_C-\mu_v(1|\omega^v,\emptyset))\Big)\end{pmatrix} + (1-\lambda)\begin{pmatrix} \rho\Big(1-G(q_C-\mu_v(1|-1,\tau_{I_d}))\Big) \\ +(1-\rho)\Big(1-G(q_C-\mu_v(1|-1,\emptyset))\Big)\end{pmatrix} \quad (8)$$

Across all treated villages with value $z_v = \tau_I$, some experience project completion (proportion $(\alpha(\tau_I)\theta)$, others see project failure. Then, the incumbents' average vote share is:

$$\begin{aligned}\mathcal{S}(\theta,\tau_I) =& \alpha(\tau_I)\theta\begin{pmatrix} \lambda\rho\Big(1-G(q_C-\mu_v(1|1,\tau_I))\Big) \\ +\lambda(1-\rho)\Big(1-G(q_C-\mu_v(1|1,\emptyset))\Big) \\ +(1-\lambda)\rho\Big(1-G(q_C-\mu_v(1|-1,\tau_I))\Big) \\ +(1-\lambda)(1-\rho)\Big(1-G(q_C-q_I)\Big)\end{pmatrix} \\ &+ (1-\alpha(\tau_I)\theta)\begin{pmatrix} \lambda\rho\Big(1-G(q_C-\mu_v(1|0,\tau_I))\Big) \\ +\lambda(1-\rho)\Big(1-G(q_C-\mu_v(1|0,\emptyset))\Big) \\ +(1-\lambda)\rho\Big(1-G(q_C-\mu_v(1|-1,\tau_I))\Big) \\ +(1-\lambda)(1-\rho)\Big(1-G(q_C-q_I)\Big)\end{pmatrix}\end{aligned} \quad (9)$$

We can then use Equation 5, Equation 7, and Equation 9 to map empirical and theoretical quantities. To do so, it is important to remember that the researchers compare the treatment and control groups for a specific value of the treatment. That is, the researchers look at the difference

in vote shares between treated villages with treatment value $z^v \in \{0,1\}$ and control villages that would have received treatment $z^v \in \{0,1\}$, *should they have been treated*.

In practice, the estimand for the performance treatment with value $Z^v = \omega^v$ corresponds to the difference between $\mathcal{S}(\theta, \omega^v)$ from Equation 7 and the average vote share in control villages that experience the same project outcome $\omega^v \in \{0,1\}$. That is, the theoretical equivalents of our empirical quantities for project failure ($\omega^v = 0$) and project completion ($\omega^v = 1$) are, respectively:

$$a_2 = \rho(1-\lambda)\Big(G(q_C - \mu_v(1|-1,0)) - G(q_C - q_I)\Big) \tag{10}$$

$$a_2 + a_3 = \rho(1-\lambda)\Big(G(q_C - \mu_v(1|-1,1)) - G(q_C - q_I)\Big) \tag{11}$$

In turn, the estimand for the corruption treatment with value $Z^v = \tau_I$ corresponds to the difference between $\mathcal{S}(\theta, \tau_I)$ from Equation 9 and the average vote share in control villages that are represented by an incumbent of type $\tau_I \in \{0,1\}$. In this case, the theoretical equivalents of our empirical quantities for corrupt incumbent ($\tau_I = 0$) and honest incumbent ($\tau_I = 1$) are, respectively:

$$\begin{aligned} a_2 =& \rho(1-\lambda)\Big((1 - G(q_C)) - (1 - G(q_C - q_I))\Big) \\ &+ \rho\lambda(\alpha(0)\theta)\Big((1 - G(q_C)) - (1 - G(q_C - \mu_v(1|1,\emptyset))\Big) \\ &+ \rho\lambda(1-\alpha(0)\theta)\Big((1 - G(q_C)) - (1 - G(q_C - \mu_v(1|0,\emptyset)))\Big) \end{aligned} \tag{12}$$

$$\begin{aligned} a_2 + a_3 =& \rho(1-\lambda)\Big((1 - G(q_C - 1)) - (1 - G(q_C - q_I))\Big) \\ &+ \rho\lambda(\alpha(1)\theta)\Big((1 - G(q_C - 1)) - (1 - G(q_C - \mu_v(1|1,\emptyset))\Big) \\ &+ \rho\lambda(1-\alpha(1)\theta)\Big((1 - G(q_C - 1)) - (1 - G(q_C - \mu_v(1|0,\emptyset)))\Big) \end{aligned} \tag{13}$$

The estimands for the performance treatment are only a function of the fundamentals of the models, i.e., the context (the proportion of informed voters, of honest types, the average probability projects are successful). The same does not hold true for the estimands of the corruption treatment. They are also a function of the *realized* environment θ. For each draw of observations, the environment θ varies and the estimand changes. In other words, the conditional difference in mean vote shares fail to permit comparability for corruption treatment.

Proposition 1. *The conditional different in mean vote shares permits comparability under the performance treatment, but not under the corruption treatment.*
Proof: All proofs are collected in Online Appendix B.

The economic environment in which the incumbents operate is the key reason behind the failure of comparability for corruption treatments when the researchers employ the conditional difference in mean vote shares. The environment θ matters because it affects the *proportion* of villages where voters, informed by Nature, observe signals $s_i^v = \omega^v = 1$ and of villages where voters observe signals $s_i^v = \omega^v = 0$. In the first category of villages, the incumbent's vote share is relatively large (since the project outcome is a signal of honesty under the assumption that $\alpha(1) > \alpha(0)$); in the second category, it is relatively low. The better the economic environment, the greater the proportion of villages with project completion, and the higher the electoral standing of the incumbent absent any sort of intervention. This makes the estimand for the corruption treatment situation-specific and, thus, always varying. Comparability fails to be achieved.

This problem does not arise for the performance treatment since the researchers focus on a group of villages with a particular project outcome (completion or failure). The environment plays no role under the assumption that the voters do not observe θ (if they did, comparability would never be permitted for any treatment). Circumstances do not affect the estimand and comparability is achieved.

Possible remedies for comparability

Having illustrated how the conditional difference in mean vote shares does not always permit comparability, we now turn to possible remedies for corruption treatments.

One solution, at first sight appealing, would be to account for the realized circumstance θ. However, as θ is drawn from a continuous probability density function, no two studies are ever conducted under the same circumstances. No two studies would then measure the same estimands, making comparability impossible to achieve with this strategy. In what follows, we argue for a somewhat more radical departure from the usual conditional difference in mean vote shares.

Our suggestion involves a change of counterfactuals. Researchers should not simply ask themselves what if the control group were to be treated (as researchers implicitly do when employing Equation 2). They should also inquire: what if the treated units had not received treatment?

Thus, instead of conditioning only on what control units would have observed if treated, researchers should also condition on what the treated villages would have learned absent treatment. We label this approach the *augmented conditional difference in mean vote shares.*

In our illustrative example, this corresponds to running the following saturated model:

$$Y_{vd} = b_0 + b_1 O_{vd} + b_2 T_{vd} + b_3 Z_{vd} + b_4 O_{vd} \times T_{vd} + b_5 O_{vd} \times Z_{vd} + b_6 T_{vd} \times Z_{vd} + b_7 O_{vd} \times T_{vd} \times Z_{vd} + \varepsilon_{vd}, \quad (14)$$

with $T_{vd} \in \{0, 1\}$ denoting the treatment status of a village v in district d, $Z_{vd} \in \{0, 1\}$ whether the incumbent is revealed to be corrupt or honest, and $O_{vd} \in \{0, 1\}$ capturing whether the village has experienced project completion ($O_{vd} = \omega^v = 1$) or failure ($O_{vd} = \omega^v = 0$)

The augmented conditional difference in mean vote shares now guarantees that treated villages and control villages are compared for the same realisation of the project outcome. For example, b_2 corresponds to the difference in vote shares between control and treated units that are represented by a corrupt incumbent *and* experienced a project failure. That is, using our theoretical model:

$$b_2 = \rho(1-\lambda)\Big((1 - G(q_C)) - (1 - G(q_C - q_I))\Big) + \rho\lambda\Big((1 - G(q_C)) - (1 - G(q_C - \mu_v(1|0, \emptyset)))\Big) \quad (15)$$

In turn, $b_2 + b_4$ measures the impact of revealing that the incumbent is corrupt in villages where the project outcome has been completed. Again, we can map this into our theoretical quantity to obtain:

$$b_2 + b_4 = \rho(1-\lambda)\Big((1 - G(q_C)) - (1 - G(q_C - q_I))\Big) + \rho\lambda\Big((1 - G(q_C)) - (1 - G(q_C - \mu_v(1|1, \emptyset)))\Big) \quad (16)$$

We can perform the same equivalence for the impact of revealing that the incumbent is honest ($b_2 + b_6$ for project failure and $b_2 + b_4 + b_7$ for project completion). In all cases, the estimands are a function of the beliefs and the proportion of treated voters, and the percentage of voters informed by Nature. None of these quantities are circumstances-specific, the environment in which the study takes place does not play a role any more. All studies estimate the same estimands. As a result,

Proposition 2. *The augmented conditional difference in mean vote shares always permits comparability for corruption treatments.*

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 variables which remove all dependence of the estimands on the environment in which the study takes place, a hard ask. We illustrate this point further in our general model below as well as in Appendix A where we use the potential outcome framework. Instead, we now turn to individual level outcomes where further solutions are available to researchers.

Individual-level outcomes

For reasons of cost, researchers often 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 the effect of the informational campaign on voting intentions, which if truthfully reported are equivalent to vote shares.

The empirical specification scholars employ then regress Y_{ivd} the reported voting intention (or vote choice) for the incumbent by individual i in village v in district d on whether an individual i is treated in village v ($T_{ivd} \in \{0,1\}$) and the value of the treatment ($Z_{vd} \in \{0,1\}$, with $Z_{vd} = 1$ if the voters learn about project completion in the performance treatment or about the honesty of the incumbent in the corruption treatment). Researchers then run the following regression (with village fixed effects, δ_v to remove village-specific attributes):

$$Y_{ivd} = \delta_v + c_0 + c_1 T_{ivd} + c_2 Z_{vd} + d_3 T_{ivd} \times Z_{vd} + \epsilon_{ivd} \tag{17}$$

The only difference between individual-level and village-level analyses is that the researchers only include compliers in the first approach, which compare voters who do receive the additional piece information versus those who did not. As such, proceeding along the same line as above

reveals that for treatment value $Z_{vd} = 0$, the estimand satisfies $c_1 = \frac{a_1}{\rho}$ and for treatment value $Z_{vd} = 1$, the estimand satisfies $c_1 + c_3 = \frac{a_1 + a_3}{\rho}$.

As a result and for the exact same reason as before, the conditional difference in mean voting intentions fail to achieve comparability for corruption treatments. The estimands is again a function of the proportion of successful project, which is circumstance specific (see Equations 12 and 13).

Proposition 3. *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 under the performance treatment, but not under the corruption treatment.*

We can then naturally extend the result of Proposition 2 to the augmented conditional difference in mean voting intentions. As before, this empirical specification can offer a potential solution to ensure comparability, but it requires that researchers account for all possible electorally relevant pieces of information villagers may receive in the control condition. With individual-level outcomes, researchers can also exploit another avenue to recover comparability. They can make use of a well-known property of voters' evaluation of their representative. Prior opinions (prior because measured before applying the treatment) about the incumbent's honesty capture all the relevant information available to villagers, without the need for researchers to make any assumptions. In other words, voters' interim belief (to use the formal language) about the likelihood the office-holder is corrupt is a sufficient statistic for all the factors that affect voters' view of their office-holder absent intervention. Conditioning on such beliefs, hence, removes all dependence on the environment. All studies measure the same estimand and comparability is restored. Further, information on voters' prior evaluations is relatively easy for researchers to collect. In our illustrative example where interim beliefs can only take one of three values (one after observing project completion, one after observing project failure, one after observing nothing), a simple three-level Likhert scale would suffice. In general, as we discuss above, a feeling thermometer is better adapted.

Label the *belief augmented difference in mean voting intentions* the difference in voting intentions when researchers condition on both the treatment and villagers' evaluation of the incumbent prior to the treatment with a three-level Likhert scale. We obtain:

Proposition 4. *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 ($\Pi_{ivd} = 2$), an average opinion ($P_{ivd} = 1$), or a bad opinion ($\Pi_{ivd} = 0$) of the office-holder. These three levels correspond, in turn, to observing project completion, observing nothing, observing project failure. The belief augmented conditional differences in mean voting intentions for corruption treatments can be recovered by running the following regression:

$$\begin{aligned} Y_{ivd} =& \delta_v + d_0 + d_1 T_{ivd} + d_2 Z_{vd} + \sum_{j=1}^{2} d_{2+j} \mathbb{I}_{\{\Pi_{ivd}=j\}} + d_5 T_{ivd} \times Z_{vd} + \sum_{j=1}^{2} d_{5+j} T_{ivd} \times \mathbb{I}_{\{\Pi_{ivd}=j\}} \\ &+ \sum_{j=1}^{2} d_{7+j} Z_{vd} \times \mathbb{I}_{\{\Pi_{ivd}=j\}} + \sum_{j=1}^{2} d_{9+j} T_{ivd} \times Z_{vd} \times \mathbb{I}_{\{\Pi_{ivd}=j\}} + \varepsilon_{ivd} \end{aligned} \tag{18}$$

In Equation 18, the reference category for interim beliefs is $\Pi_{ivd} = 0$ (so the worst possible opinion of the incumbent prior to treatment). The coefficient d_2 then captures the impact of revealing voters with low interim belief that the incumbent is corrupt. The sum $d_2 + d_5 + d_6 + d_8 + d_1 0$ corresponds to the effect of revealing that the incumbent is honest ($Z_{vd} = 1$) to voters who have intermediate opinion of the office-holder to begin with (the voters $\Pi_{ivd} = 1$). Other estimands can be recovered in a similar fashion.

Two aspects of our approach are important to note. The first is that we use dummies for each different beliefs. Indeed, interacting the treatment status (T_{ivd}) or treatment value (Z_{vd}) with prior opinions imposes a linear effect of the treatment for different levels of beliefs. In our setting, this would be equivalent of assuming that the function $G(\cdot)$ is linear. There is no reason for it to be so and a flexible approach with dummies is more appropriate to avoid the risk of model misspecification. Second, there are as many estimands as combinations of treatment value - interim belief. So with a binary treatment value and a three-level scale, we have $2 \times 3 = 6$ estimands. This soon becomes impracticable as a solution when the number of possible values for the interim beliefs and/or for the treatment values become large. When we turn to our general model in the next section, we discuss an alternative solution. Before that, we want to take advantage of our illustrative example

to highlight the difference between our approach described in Equation 18 and the specifications adopted in an important recent scholarly endeavour to accumulate knowledge, Metaketa I (2020).

Like Dunning et al. (2020), our solution makes use of voters' beliefs prior to the researchers' intervention. We part ways with this work when it comes to how researchers should employ this variable. Voters' prior evaluations of office-holders in the Metaketa I studies serve to obtain a consistent definition of 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? 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. It is just a way to rewrite the value of the treatment Z_{vd}. It does nothing to remove the potential effect of the circumstances in which the interventions take place. Each study still yields an estimate of a circumstance-specific estimand.

To see this, let's take the corruption treatment when voters are informed that the incumbent is honest. This is good news for all voters (all treated voters improve their posterior relative to the prior). Hence, conditioning on this good news is the same as conditioning on treatment value $Z_{vd} = 1$; the Metaketa I specification measures exactly the same estimand as in Equation 2. And this approach, we know, does not permit comparability (we recognize that the information provided by Metaketa I is different than ours, but the spirit of the idea is the same). In contrast, our approach does not define good or bad news relative to voters' initial evaluation of the incumbent. Indeed, it does not even rely on determining what good news is (though we could to generate predictions about the ranking of the estimands). Rather, it proposes to use voters' priors to identify stable control groups across studies conducted in similar contexts, thus ensuring comparability.

3 The General model

In this section, we illustrate how the results from our illustrative example carry over to a more general model. The model we study here is more abstract by design, as it is meant to illustrate that

the difficulties of accumulating knowledge are not model-dependent. We first describe the general set-up, before illustrating the issues for comparability and the possible solutions.

Set-up

We maintain the geographical set-up from the previous section, and study a two-period game. The country researchers intervene in is divided into a mass D of districts. Each district $d \in D$ is represented in period 1 by an incumbent (I_d), which may be replaced by a challenger (C_d) in the election at the end of period 1. In each district, there is a mass V of villages and in each village v, there is a mass 1 of voters.

We now assume that the type $\widetilde{\tau_J}$ of a political $J \in \{I, C\}$ is distributed according to the cumulative distribution function (CDF) $Q_J(\cdot)$ and probability distribution function (pdf) $q_J(\cdot)$ over the real line. We retain the assumption that the distribution of types is the same in each district. A voter i in village v does not observe the type of any politician. She, however, receives an informative signal s_i^v. The distribution of the signal is affected by the realisation of the incumbent's type in the district (τ_{I_d}), and the environment in which the incumbent operates, which we still denote θ. Formally, each voter's signal $\widetilde{s_i^v}$ is distributed according to the CDF $M(\cdot|\tau_{I_d}, \theta)$ and pdf $m(\cdot|\tau_{I_d}, \theta)$. The actual environment θ is unknown to voters. However, it is common knowledge that this state of the world is distributed according to the CDF $F(\cdot)$ and pdf $f(\cdot)$ over the real line, with $f(\cdot)$ strictly positive over an interval of non-measure zero. Notice that while we make no assumption on the signal generating process, so the distribution $M(\cdot|\cdot)$ can very well be endogenous to the incumbent and/or challenger's actions.

Voters' second-period payoff from electing politician $J \in \{I_d, C_d\}$, which is the only one that enters our computations, is $V(\tau_J) + \sigma_i^J$. $V(\tau)$ is a strictly increasing function in τ so that voters prefer high values of a politician's type to low values. In addition, a voter's evaluation of politicians is affected by idiosyncratic shocks σ_i^J.

An informational campaign, in turn, takes the form of the researchers randomly selecting a subset of villages and dividing them into two groups, control and treatment. In treated villages, the researchers provide a proportion ρ of voters with an additional signal z^v. This signal is distributed according to the CDF $\Psi(\cdot|\tau_{I_d}, \theta)$ and pdf $\psi(\cdot|\tau_{I_d}, \theta)$ before the election. Note that all treated voters

in a village receives the same piece of information; voters from two distinct villages may be exposed to different treatment values (i.e., $z^v \neq z^{v'}$).

The timing is similar to the illustrative example: 1. In period 1, Nature draws the incumbents' and the challengers' types in all districts. 2. Nature draws θ and all the s_i^v observed by voter i in village v. 3. Researchers randomly select some villages and, among the treated samples, ρ voters receive the informational treatment. 4. The voters observe their idiosyncratic shock σ_i^J ($J \in \{C_d, I_d\}$) and cast their ballot. 5. The politician who receives the most votes in a district is elected. 6. The game moves to period 2 and ends with payoffs being realized.

Remarks on the set-up

Throughout the analysis, we distinguish between random variables, denoted by $\widetilde{\cdot}$, and the realisation of a random variable, without the tilde accent. As in our simple example, this distinction proves critical. Characterised by their CDF and pdf, random variables are context-specific (they may vary from countries to countries). As we fix the context, random variables remain unchanged from one study to the next in our paper. The realization of a random variable is, in contrast, circumstance-specific. Two teams of researchers who intervene at two different points in time *within the same context* would face different circumstances (different types of office-holders, different economic environments). An empirical estimation which yields an estimate of a circumstance-dependent estimand does not permit comparability as per our definition (Definition 1).

To avoid dealing with corner solutions, which complicate the analysis and the notation without adding any insight, we make a few assumptions. We normalize the utility functions so that $\lim_{\tau \to -\infty} V(\tau) = -1$ and $\lim_{\tau \to \infty} V(\tau) = 1$. The idiosyncratic shocks satisfy $\sigma_i^I - \sigma_i^C$ is continuously distributed according to the CDF $G(\cdot)$, satisfying $G(-2) > 0$ and $G(2) < 1$. To determine a vote share for each village, we assume that for each signal s_i^v, the full distribution of $\sigma_i^I - \sigma_i^C$ is realized. This is guaranteed if the set of possible signals with strictly positive probability of realisation is discrete. This is a more stringent assumption on the relative masses of signals and shocks if $M(\cdot|\cdot)$ is continuous. In this case, when the assumption is relaxed, we can only compute an expected vote share in each village. The expected vote share would, however, correspond to the average of vote shares across a mass of villages, which, we assume, researchers have access to. Hence, the assumption is without loss of generality. In a more important way, we also impose that there is a mass of

villages for each possible realisation of the informational campaign z^v and each possible realisation of the incumbent's type τ_{I_d}. This actually corresponds to a best case scenario for researchers (as it eliminates all sources of statistical noise). This assumption always holds in our setting when the set of possible signals provided by researchers is discrete, which is almost always the case in the empirical literature, or when the type of the incumbent can only take a limited number of values, as is often, though not always, assumed.

Regarding the treatment, as in our illustrative example, the assumption that some voters in a treated village do not receive the treatment has two implications: (i) the effect of the treatment can be measured at the village level or at the individual level and (ii) at the village level, the researchers can only recover an intention to treat (unless they know ρ). Of course, as we noted before, individual-level analyses require researchers to randomize who gets the treatment within each treated village.

Finally, for illustrative purposes, it is useful to relate our general model to our motivating example. For the politicians' types, our motivating example imposes $q_J(0) = 1 - q_J$ and $q_J(1) = q_J$ (since we have made no assumption on the distribution functions, we can impose that they have mass points). To see how the signals from the illustrative example are a special case of our general set-up, notice first that we have not imposed any i.i.d. assumption. We can partition villages into two subsets $\mathcal{V}_0$ and $\mathcal{V}_1$. The probability that a village v belongs to $\mathcal{V}_1$ is $\alpha(1)\theta$. Voters do not directly observe the subset their village belong to. Conditional on a village being in the subset $\mathcal{V}_1$ (on top of the incumbent's type and realised environment), the distribution of signals is i.i.d. and satisfies $m(-1|\tau_{I_d}, \theta^1, v \in \mathcal{V}_1) = 1 - \lambda$ and $m(1|\tau_{I_d}, \theta^1, v \in \mathcal{V}_1) = \lambda$. Conditional on a village being in the subset $\mathcal{V}_0$ (on top of the incumbent's type and realised environment), the distribution of signals is i.i.d. and satisfies $m(-1|\tau_{I_d}, \theta^1, v \in \mathcal{V}_0) = 1 - \lambda$ and $m(0|\tau_{I_d}, \theta, v \in \mathcal{V}_1) = \lambda$. In the case of corruption treatment, the signal $\widetilde{z^v}$ is distributed according to $\psi(z^v = \tau_{I_d}|\tau_{I_d}, \theta) = 1$. In the case of performance treatment, the signal $\widetilde{z^v}$ is distributed according to $\psi(z^v = \omega|\tau_{I_d}, \theta^1, v \in \mathcal{V}_\omega) = 1$ for $\omega \in \{0, 1\}$. We can then define the other distributions and the utility function accordingly to complete the mapping.

Analysis

In what follows, we denote the posterior of a voter from a village in the control group who has received a signal s_i^v, which we denote $\mu(\tau_{I_d}|s_i^v, \emptyset)$. Similarly, a voter in a treated village observes both s_i^v and the informational treatment z^v, and we denote her posterior $\mu(\tau_{I_d}|s_i^v, z^v)$. The exact formula for the posteriors are provided in Online Appendix B.2.

Proceeding very much along the same lines as in our illustrative example (see Online Appendix B.2 for more details), we recover the following vote share for the incumbent in the control group (Equation 19) and in the treated group (Equation 20), respectively.

As the control group consists of a mass of villages with a mass of inhabitants in each, the average vote share is a weighted average of individual votes across all possible realizations of the signal $\widetilde{s}_i^v$ and across all realization of the incumbent's type $\widetilde{\tau}_I$. Unlike our illustrative example, since types and signals are possibly continuously distributed, we need to use the integral symbol, which serves as a shorthand for integration over the real line, to capture the value of the different weights.

$$\mathcal{S}(\theta, \emptyset) = 1 - \int\int \overbrace{G\left(\int V(\widetilde{\tau}_I)\mu(\widetilde{\tau}_I|\widetilde{s}_i^v, \emptyset)d\widetilde{\tau}_I - \int V(\widetilde{\tau}_C)q_C(\widetilde{\tau}_C)d\widetilde{\tau}_C\right)}^{\text{Vote choice of } i \text{ after signal } s_i^v} \underbrace{dM(\widetilde{s}_i^v|\widetilde{\tau}_I, \theta)}_{\text{Weight: signal}} \underbrace{dQ_I(\widetilde{\tau}_I)}_{\text{Weight: type}} \tag{19}$$

The average vote share in the treated group takes a similar form with the addition of some treated voters (proportion ρ of voters in treated units):

$$\begin{aligned}\mathcal{S}(\theta^1, z^v) = 1 &- \rho \int\int \overbrace{G\left(\int V(\widetilde{\tau}_I)\mu(\widetilde{\tau}_I|\widetilde{s}_i^v, z^v)d\widetilde{\tau}_I - \int V(\widetilde{\tau}_C)q_C(\widetilde{\tau}_C)d\widetilde{\tau}_C\right)}^{\text{Vote choice of } i \text{ after signal } s_i^V \text{ and treatment } z^v} \underbrace{dM(\widetilde{s}_i^v|\widetilde{\tau}_I, \theta, z^v)}_{\text{Weight: signal}} \underbrace{dQ_I(\widetilde{\tau}_I|z^v)}_{\text{Weight: type}} \\ &- (1-\rho) \int\int \overbrace{G\left(\int V(\widetilde{\tau}_I)\mu(\widetilde{\tau}_I|\widetilde{s}_i^v, \emptyset)d\widetilde{\tau}_I - \int V(\widetilde{\tau}_C)q_C(\widetilde{\tau}_C)d\widetilde{\tau}_C\right)}^{\text{Vote choice of } i \text{ after signal } s_i^v} \underbrace{dM(\widetilde{s}_i^v|\widetilde{\tau}_I, z^v)}_{\text{Weight: signal}} \underbrace{dQ_I(\widetilde{\tau}_I|z^v)}_{\text{Weight: type}}\end{aligned} \tag{20}$$

Notice that, in treated villages, the distribution of signals and types may be a function of information provided the treatment (e.g., in our motivating example, upon receiving the corruption treatment, only good or bad type would be in the researchers' sample). Hence, the distributions of

$\widetilde{s}_i^v$ and $\widetilde{\tau}_I$ are conditional on z^v, whether the voters are compliers (probability ρ) or not (probability $1-\rho$).

We are now ready again to study the estimands measured with a conditional difference in mean vote shares. As we noted in our illustrative example, for a treatment value z^v, this corresponds to comparing $\mathcal{S}(\theta, z^d)$ with the average vote share in control villages that would have received treatment value z^v if treated. The conditional difference in mean for treatment value z^v measures $\mathcal{S}(\theta^1, z^v) - E\left(S_v(\tau_{I_d}, \theta^1, \emptyset)|\widetilde{z} = z^v\right)$ and equals:

$$\rho\Bigg(\int\int G\left(\int V(\widetilde{\tau}_I)\mu(\widetilde{\tau}_I|\widetilde{s}_i^v, \emptyset)d\widetilde{\tau}_I - \int V(\widetilde{\tau}_C)q_C(\widetilde{\tau}_C)d\widetilde{\tau}_C\right) dM(\widetilde{s}_i^v|\widetilde{\tau}_I, \theta, z^v)dQ_I(\widetilde{\tau}_I|z^v)$$
$$- \int\int G\left(\int V(\widetilde{\tau}_I)\mu(\widetilde{\tau}_I|\widetilde{s}_i^v, z^v)d\widetilde{\tau}_I - \int V(\widetilde{\tau}_C)q_C(\widetilde{\tau}_C)d\widetilde{\tau}_C\right) dM(\widetilde{s}_i^v|\widetilde{\tau}_I, \theta, z^v)dQ_I(\widetilde{\tau}_I|z^v)\Bigg) \quad (21)$$

The key difference between the control and treatment units is that the researchers' intervention z^v affects the posterior in the case of the treatment (last line) of Equation 21, whereas it only affects the possible distributions of signals (given their village-level correlation) in the control group.

We can easily adapt the analysis, as we have done above, to recover the estimand for individual level analysis. If the researchers randomize within villages, then the treatment effect is, like in our illustrative example, the ATE.

$$\int\int G\left(\int V(\widetilde{\tau}_I)\mu(\widetilde{\tau}_I|\widetilde{s}_i^v, \emptyset)d\widetilde{\tau}_I - \int V(\widetilde{\tau}_C)q_C(\widetilde{\tau}_C)d\widetilde{\tau}_C\right) dM(\widetilde{s}_i^v|\widetilde{\tau}_I, \theta, z^v)dQ_I(\widetilde{\tau}_I|z^v)$$
$$- \int\int G\left(\int V(\widetilde{\tau}_I)\mu(\widetilde{\tau}_I|\widetilde{s}_i^v, z^v)d\widetilde{\tau}_I - \int V(\widetilde{\tau}_C)q_C(\widetilde{\tau}_C)d\widetilde{\tau}_C\right) dM(\widetilde{s}_i^v|\widetilde{\tau}_I, \theta, z^v)dQ_I(\widetilde{\tau}_I|z^v) \quad (22)$$

As direct observation from Equation 21 and Equation 22 indicates, the estimands from the conditional difference in mean vote shares can be a function of θ via the distribution of signals $M(\widetilde{s}_i^v|\widetilde{\tau}_I, \theta, z^v)$. As a result, we obtain the following proposition, which generalizes Proposition 1.

Proposition 5. *The conditional difference in means permits comparability for informational treatment z^v at the village level or at the individual level if and only if $M(\cdot|\tau_I, \theta, z^d) = M(\cdot|\tau_I, z^d)$ for all τ_I satisfying $q_I(\tau_I) > 0$ and all θ satisfying $f(\theta) > 0$.*

The conditional difference in means permits comparability at the village level or at the individual level

for all treatment values if and only if $M(\cdot|\tau_I, \theta, z^d) = M(\cdot|\tau_I, z^d)$ *for all* z^v *satisfying* $\psi(z^v|\tau_I, \theta) > 0$ *for all* τ_I *satisfying* $q_I(\tau_I) > 0$ *and all* θ *satisfying* $f(\theta) > 0$.

The proposition makes two points. First, from a statistical perspective, the conditional difference in mean vote shares permits comparability if and only if z^v is a sufficient statistic for θ when it comes to the voters' signals s_i^v (i.e., once we condition on z^v, then the distribution of voters' signals s_i^v does not depend on θ). In maybe more intuitive terms, the typically employed conditional difference in mean vote shares permits comparability if and only if conditioning on the treatment removes all dependence of the incumbent's vote share to the environment in which the researchers intervene. Notice that this condition is not testable within a single intervention. It would require to measure the vote shares conditional on z^v under different circumstances. This requires multiple interventions *at different times*.

Remedies for recovering comparability and their feasibility

Like for our illustrative example, researchers cannot turn to conditioning on circumstances to recover comparability. As the environment is drawn from a continuous distributions, no two studies are ever taking place in the same circumstance. An alternative solution is to find a variable which may serve as sufficient statistics for the environment, such as the project outcome in our illustrative example. Researchers may then recover comparability by conditioning, not controlling on the different values of this variable (as the circumstances *interact* with the treatment via the voters' beliefs). Whether such variable exists is a difficult question to answer, though.

Proposition 6. *For village level analysis, the augment conditional difference in means, which conditions on a sufficient statistics for the realized circumstances, permits comparability.*

Researchers engaging in individual-level analyses, especially employing experiments within surveys, have another more promising avenue for recovering comparability. This solution corresponds once more to a change of counterfactual. Rather than simply asking what would have happened to the control voters if treated, researchers need to also wonder "and what would have been treated voters' opinion of the incumbent if not treated." That is, as we noted in our illustrative example, researchers can condition on voters' interim beliefs about τ_I to restore comparability.

Proposition 7. *For individual level analyses, the belief-augmented conditional difference in mean vote shares permits comparability.*

Our solution again makes use of the properties of interim beliefs (those measured at the beginning of survey of voters, prior to treatment if any). Those beliefs represent a sufficient statistics for all the electorally relevant information voters have access to absent treatment . It removes dependence on the signal, its distribution, and, thus, the realized environment θ as long as the measured beliefs are related to the treatment used. As the estimands are no longer a function of the circumstances, researchers recover comparability.

Two important aspects of our recommended solution are worth stressing. First, the voter's interim beliefs *interacts* with the treatment (see Online Appendix B.2 for details). Thus, simply controlling (rather than conditioning) for a voter's opinion of the incumbent is not sufficient to recover comparability. Second, it is critical to condition on a fine grained measure of the voters' beliefs, such as thermometer feeling or grades over a large scale. The cost of not doing so is to reintroduce the environment into the estimands. Indeed, suppose that the researchers employ a four-scale categories to measure belief: S_1, S_2, S_3, S_4. There then exists four set of signals—$[s_0, s_1]$,$[s_1, s_2]$,$[s_2, s_3]$,$[s_3, s_4]$, with s_0 and s_4 the bounds on the signal space—such that if $s_i^v \in [s_{\chi-1}, s_\chi]$, then the surveyed voter respond S_χ for all $\chi \in \{1, 2, 3, 4\}$. This means that the vote share (or voting intention) of the incumbent conditional on S_χ, $\mathcal{S}(\theta^1, z^v|S_\chi)$ is equal to:

$$1 - \int \int_{s_{\chi-1}}^{s_\chi} G\left(\int V(\widetilde{\tau}_I)\mu(\widetilde{\tau}_I|\widetilde{s}_i^v, \iota^v)d\widetilde{\tau}_I - \int V(\widetilde{\tau}_C)q_C(\widetilde{\tau}_C)d\widetilde{\tau}_C\right) dM(\widetilde{s}_i^v|\widetilde{\tau}_I, \theta, \iota^v)dQ_I(\widetilde{\tau}_I|\widetilde{s}_i^v \in [s_1, s_2], \iota^v)$$

Unless the distribution of signals in each interval $[s_{\chi-1}, s_\chi]$ is independent of θ (in words, the economic environment affects how many voters are in each interval, but not how signals are distributed within each interval, a non-testable assumption), conditioning on broad categories of belief does not permit comparability.

As a final note, we recognize that conditioning on a fine grained measure of beliefs may prove too taxing in term of data requirements (e.g., for an one-hundred scale feeling thermometer, the numbers of variables and, thus, estimands equal one hundred times the number of treatment values). Researchers, however, can make use of a property of posteriors: posteriors are continuous in the values of interim beliefs, though not necessarily linearly (see Online Appendix B.2). Hence, a

potentially attractive solution would be to use a flexible polynomial approach (conditioning on a polynomial of interim beliefs) in individual level studies as a mean to recover comparability. Results could then be presented in the form of plots with voters' prior opinion on the x-axis and estimates on the y-axis.

4 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.

Reaching unbiasedness and obtaining comparability are complementary objectives. The first 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 essential for the production of knowledge in social sciences.

Using informational campaigns and their impact on electoral outcomes as example, we highlight that comparability should not be assumed. The conditional difference in mean vote shares, 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).

Our paper suggests that comparability is much more a concern than previously thought even when external validity is not an issue. Comparability, however, is not an impossible goal within a particular context. We offer several recommendation to recover comparability. They are all derived from the same underlying principle: a change of counterfactual. Researchers should not just ask themselves what the control units would have learned if treated (as the conditional difference in

mean vote shares implies), they should also wonder what the treated units would have learned if not treated (what the augmented conditional difference in means guarantees, as we argue). We offer a practical solution for randomization performed at the level of voters within villages: to flexibly condition on voters' opinions of the incumbent at the beginning of surveys. In the spirit of the theoretical implications of empirical models, we hope that our work with its negative and positive results highlights how theory and empirics can fruitfully be joined for the production and accumulation of knowledge.

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. 2021. "Able and Mostly Willing: An Empirical Anatomy of Information's Effect on Voter-Driven Accountability in Senegal." American Journal of Political Science, forthcoming. Available at https://doi.org/10.1111/ajps.12591.

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. 2022. " Candidate entry and vote choice in the wake of incumbent performance transparency initiatives." Working Paper. Available at https://osf.io/qwcek/.

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

Incerti, Trevor. 2020. "Corruption information and vote share: A meta-analysis and lessons for experimental design." American Political Science Review 114(3): 761-774.

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.

Slough, Tara and Scott A. Tyson. 2021. "External Validity and Meta-Analysis." Working Paper. Available at https://drive.google.com/file/d/11G02RN6sOlGJtPaEyCjQlAYE9XL7GNnQ/view.

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