Metaketa I:

Information, Accountability, and Cumulative Learning

Thad Dunning, Guy Grossman, Macartan Humphreys, Susan D. Hyde, and Craig McIntosh, eds.

with

Claire Adida, Eric Arias, Taylor C. Boas, Mark Buntaine, Sarah Bush, Simon Chauchard, Anirvan Chowdhury, Jessica Gottlieb, F. Daniel Hidalgo, Marcus Holmlund, Ryan Jablonski, Eric Kramon, Horacio A. Larreguy, Malte Lierl, John Marshall, Gwyneth McClendon, Marcus A. Melo, Gareth Nellis, Daniel Nielson, Paula Pickering, Melina Platas, Pablo Querubín, Pia Raffler, Catlan Reardon, and Neelanjan Sircar

This draft: February 5, 2018

Chapters 2 and 11 — Not for circulation outside of the EGAP Vanderbilt meetings

Contents

Chapter 2

The Metaketa Initiative

THAD DUNNING, GUY GROSSMAN, MACARTAN HUMPHREYS, SUSAN D. HYDE, AND CRAIG MCINTOSH

2.1 The Challenge of Cumulative Learning

Researchers, practitioners, and policymakers often share an important goal: they want to use research to understand how the world works and to assess what interventions, policies, or programs can make things better. Research, many hope, can guide policy choices in new situations.

Yet, such insights are difficult to acquire from any one research study. In the social and political domain, unlike in some natural science domains, immutable laws that hold across time and place may be the exception rather than the rule. Social scientists often point to the limited "external validity" of particular studies—that is, the extent to which the findings from one study may travel to other interventions, contexts, and study populations.[1] The specifics and vagaries of implementation of particular studies can also

[1] Campbell and Stanley (1966) define external validity in terms of "*generalizability*: To what populations, settings, treatment variables, and measurement variables can [an] effect be generalized?"

make the generalizability of findings uncertain.[2] Therefore, results from a single study may provide an unconvincing basis for asserting generalities, and only a tentative basis for extrapolation. At the same time, some findings may in fact generalize beyond the context of a single research study. The extent of external validity should thus ideally be evaluated with evidence, rather than being left to conjecture.[3] In other words, the common effects of similar interventions need to be assessed and possibly demonstrated, rather than assumed or rejected a priori.

One sensible solution to the challenge of generalizability, it appears, is to combine the results of multiple studies on the same topic. Such "meta-analysis" is commonplace in some physical sciences, and it appears from time to time in the social sciences, too.[4] Aggregating evidence from multiple studies may give us greater purchase on whether—and even why—results differ across distinct contexts, or whether instead findings point us in the same direction across settings.

Unfortunately, meta-analysis as a solution to the problem of generalizability faces several critical difficulties in the social sciences. Consider in more detail the several obstacles noted in Chapter 1:

Study sparsity. Meta-analyses require as inputs several related studies on a single topic. Yet, one important challenge arises from the practice of social science and the career incentives that many scholars face. Academic research is a decentralized operation; researchers typically pursue questions that are interesting to them, and in consequence may pay less attention to knowledge accumulation over time. Moreover, there is a substantial premium placed on novelty, and relatively fewer rewards for reproducing or

[2] Berge et al. (2012); Bold et al. (2016).

[3] Campbell and Stanley (1966).

[4] See Gerber and Green (2012, Chapter 11) for several examples.

validating prior results across time and space. Scholars may benefit more professionally from publishing "groundbreaking" work, compared to the rewards of publishing work that replicates existing findings. Such career incentives to "plant the flag" imply that once a hypothesis or finding is published, too few scholars are willing to invest in corroborating that result. Thus, despite widespread acknowledgement that replication is an essential part of a research agenda, important studies are rarely replicated. This failure, contributing to study sparsity, complicates the goal of combining the results of different studies using meta-analysis.[5]

Study Heterogeneity. Formal meta-analyses also typically assume that both the intervention under consideration and the manner in which it is deployed and analyzed is constant across the different studies that they take as inputs.[6] These assumptions are necessary if researchers are to treat different studies as though they are part of the same grand study. This assumption can be a strong one even within a single study. For example, experimental researchers assume that all subjects assigned to a treatment condition receive the "same" treatment, but even respondents visited by the same canvassers in a get-out-the-vote experiment might experience a quite different treatment (depending, say, on the canvasser's mood). However, given the varied contexts in which social science studies take place–and also the incentives to differentiate studies just mentioned—this assumption is often especially strong when combining the results of different studies. The difficulty of study heterogeneity can vary across research topics: in testing the effects of, say, financial inducements to purchase malarial beds, or micro-finance schemes to jump-start small businesses, it might be plausible to assume that interventions are sufficiently similar across studies (though even there, one might ask whether subjects interpret the meaning

[5] Dunning and Hyde (2014).

[6] Gerber and Green (2012, 351).

of a financial inducement differently in different contexts). But for many interventions in the social sciences, especially in the governance space, such homogeneity cannot be assumed.

Selective reporting. Finally, related difficulties arise from the way that research is typically reported. Publication bias—the tendency of academic journals and presses to publish statistically significant (positive or negative) estimated effects, but not null results—poses another genuine threat to the validity of inferences from bodies of research. The evidence of publication bias is now quite extensive and very convincing.[7] This bias can occur not only because referees and editors fail to publish studies reporting null effects, but also because authors sometimes do not write up such results in the first place.[8] This can be a problem for the reliability of individual studies, especially to the extent that authors of individual studies engage in specification searches, data mining or "*p*-hacking" to turn up and report only significant effects.[9] Yet, publication bias also limits cumulative learning from bodies of research on a particular topic. A null effect is not a null finding; and publishing only non-null effects exaggerate our sense of the causal efficacy of particular policies or interventions.

A final important difficulty stems from **private data**. Without publicly available data, third-party researchers cannot reconstruct results to verify that authors used best practices when analyzing their raw unprocessed data, nor can they use the data to conduct systematic meta-analyses.[10] The situation has improved since scholars such as Gary

[7] See Gerber, Green and Nickerson (2001); Gerber, Malhotra et al. (2008); Simonsohn, Nelson and Simmons (2014)

[8] Franco, Malhotra and Simonovits (2014).

[9] Humphreys, Sanchez de la Sierra and van der Windt (2013); Laitin (2013).

[10] King (1995). To distinguish such efforts from the effort to replicate results with new data from new study sites, this sort of exercise is variably referred to as "internal replication," "pure replication" (Hamermesh, 2007), or a "reproduction test" (Clemens, 2017).

King drew attention to the importance of public data in the 1990s.[11] Yet while several leading journals in political science currently require data posting and even third-party verification of the match between data, code, and reported results, a recent survey of replication policies at 120 peer-reviewed political science journals found that only 19 even had a replication policy.[12] When data are private, third parties cannot readily assess the reliability of conclusions drawn from in any particular study. Mistakes are all too easy to make in a long and often complex research process; and when data are available, the record is not encouraging.[13] The inability to access data therefore remains a basic but substantial barrier to open and reliable science, as well as to aggregation.

In sum, study sparsity, study heterogeneity, selective reporting, and private data threaten the feasibility of meta-analysis—and of cumulative learning—in the social sciences. The tendency of social-scientific inferences and policy recommendations to be drawn from a single or small number of high-visibility published studies showing important treatment effects—while less visible studies that suggest null effects are not published, and implementation failures are not reported—leads to a distorted view of the likely effects of interventions. Several scholars may conduct studies on related topics; yet differences in interventions, outcomes, measurement of inputs and outputs and other aspects of study design can limit the comparability of results from such studies. Scholars' incentives to distinguish their work from previous research in the area—reflecting the returns to novelty—further undermines the effort to base conclusions on several studies of

[11] King (1995).

[12] Gherghina and Katsanidou (2013). The *APSR* currently requires posting of data and code, while the *American Journal of Political Science* recently began to require third-party verification.

[13] For instance, one attempt at replicating research—in a journal with a policy of mandatory data archiving!—found that data for only 69 of 193 articles were in fact archived; and only 58 had both data and code present. The authors could only reproduce results for 14 of 62 studies (or 23 percent) that they sought to verify (McCullough, McGeary and Harrison, 2006, 1101, 1105).

a phenomenon, rather than just one.

At issue here is how much and how reliably researchers can learn not just from a single study, but from a collection of studies on a given topic. It can be difficult to walk away from a literature on a topic with a clear understanding either of an "average" effect of a policy or program, or a clear sense of the conditions under which a particular effect may hold. These challenges underscore the importance of building strategies that allow us to better validate and aggregate findings as well as understand heterogeneous results in different contexts.

2.1.1 Internal vs. External Validity? The Rise of RCTs

These difficulties in aggregation have been underscored—and perhaps even exacerbated—by the dramatic recent growth in the use of Randomized Controlled Trials (RCTs) in the social sciences.[14]

Experiments provide a highly valuable method for understanding which policies and programs may improve socially desirable outcomes. Causal relationships in the real world (outside of controlled laboratory settings) are often obscured by confounding factors. For example, those exposed to a program under consideration may not be comparable to those who are not, in a myriad of ways that are hard to measure or observe. As the mantra goes: correlation is not causation. For this reason, RCTs are sometimes referred to as the "gold standard" of research design. Due to random assignment to intervention (or "treatment") conditions, experiments help to avoid the problem of confounding—arguably the most important of threats to the internal validity of studies.[15] Strong research design

[14] On the growth of RCTs and related methods, see McDermott (2002), Druckman et al. (2006), De Rooij, Green and Gerber (2009), Humphreys and Weinstein (2009), Hutchings and Jardina (2009), Palfrey (2009), Angrist and Pischke (2010), Dunning (2012), and Hyde (2015).

[15] Internal validity refers to the ability of a study to describe cause and effect relations in a particular

is often a first, necessary step towards reliable cumulative knowledge about cause and effect.

We therefore welcome this turn towards experiments. Yet our sense, shared with many scholars, is that to realize the full gains of this turn towards randomized designs, this focus must be complemented by other practices and institutions.[16] Moreover, the inferential advances brought about by RCTs do not in themselves solve the external validity problems noted above. Indeed, because RCTs require intense focus on design details—the implementation of which is often specific to particular contexts—RCTs sometimes lead researchers to drill down into the particularities of distinct study sites. While for many reasons we applaud this close engagement with particular settings, from the standpoint of generalizability, this could risk making matters worse. This concern is therefore sometimes cast as a tension between internal and external validity.[17]

While we note that there is no fundamental reason why a gain on internal validity means a loss on external validity, RCTs alone may do surprisingly little to facilitate cumulative learning. Consider the following reasons for concern:

Excludability. The most obvious is that experiments, by definition, involve intervention, and so claims to external validity rest on an assumption that the intervention itself does not produce effects different to what would be produced by naturalistic variation. This is sometimes referred to as the need to satisfy an exclusion restriction (Green and Gerber 2013) as well as construct validity (Morton and Williams, 2010). Excludability always poses a risk with experimentation—though plausibly it is less a concern for field

setting. Campbell and Stanley (1966) describe internal validity, in the context of experiments, as "the basic minimum without which any experiment is not interpretable: Did in fact the experimental treatments make a difference in this specific experimental instance?" See footnote 1 in this chapter for the contrast with external validity.

[16] Humphreys, Sanchez de la Sierra and van der Windt (2013); Dunning (2016).

[17] Deaton (2010).

experiments than for other forms of experimentation.

Mechanisms. The generalizability of experiments depends on understanding *why* an intervention if effective, yet experimental designs often ignore or are unable to answer "why" questions of this form.[18] Indeed, very commonly the "estimand" of an experimental research project is the average treatment effect in a population–which can be calculated with only a thin model of the way that outcomes are generated.[19] This shift in focus away from models of data generation connects to the external validity of experiments because understanding the operative mechanism can help answer the question of whether a similar intervention would work in a different context—where different mechanisms may be operative.[20] Experimental design can certainly help shed light on mechanisms.[21] Yet, features of the broader context in which an intervention takes place may not be possible to manipulate, and can clearly condition effects in ways that experiments per se do not necessarily illuminate.

Scale. Experiments are often implemented on a population that differs in scale from the population for which inferences are sought. Scholars such as Daron Acemoglu have emphasized that estimated treatment effects may not provide a reliable guide to what would happen if interventions were taken "to scale." Specifically, experiments tend to evaluate the impact of interventions in partial equilibrium, i.e., without taking account of likely reactions of critical stakeholders. Yet, general equilibrium or feedback effects would be more critical if a treatment were applied everywhere and not just to a small subset of

[18] Deaton and Cartwright (2016).

[19] For example, each unit has a potential outcome under treatment and a potential outcome under control, and which outcome is realized depends on the a unit's treatment assignment. This is a model of the data-generating process but does not stipulate a theory of any particular unit's response.

[20] See the writings of Nancy Cartwright on this point, e.g. Cartwright and Hardie (2012).

[21] For a discussion, see Gerber and Green (2012) on "implicit mediation analysis."

a population.[22]

These concerns are important for producers and users of social-science knowledge alike. The goal of much research is not simply to estimate the impact of some intervention, but to understand why the intervention had the effect that it did; to assess what would happen if the intervention were implemented in a different way or in a different context; and to add to basic social-science knowledge about where particular causal relationships obtain. To the extent that experimental research falls short of these goals—due to violation of the exclusion restriction, failure to specify mechanisms or because results cannot be scaled—it is less illuminating and less useful for both researchers and policy makers.[23]

How, then, can experimental research—with its well-understood internal-validity advantages for assessing the causal impact of interventions—address such external-validity concerns?

Among proponents of RCTs, one answer has been to argue that experiments should be replicated, in different contexts and at different scales. Thus, even if one experiment alone cannot shed light on the external validity of the answer to a research question—so the thinking goes—many such experiments can. For example, Abhijit Banerjee and Esther Duflo note that to address "concerns about generalization, actual replication studies need to be carried out. Additional experiments need to be conducted in different locations,

[22] See, e.g., Acemoglu (2010). In a small-scale field experiment, Grossman, Humphreys and Sacramone-Lutz (2016) showed that the use of SMS-based messaging services to communicate with politicians can lead to significant "flattening" of political access; however, they find no evidence for such flattening in a larger-scale national experiment.

[23] These critiques of experiments are valid and important, though we note that they are often unduly generous to observational research—which may not have the advantage sometimes claimed for it on the dimension of external validity. See e.g. Aronow and Samii (2016). The question of mechanisms in particular turns out to pose more fundamental challenges than causal inference and the challenge in studying mechanisms is no less, and perhaps greater, for observational research (Green, Ha and Bullock, 2010).

with different teams."[24] In principle, such replication should be especially feasible with experiments: to a much greater extent than is possible for many observational studies, researchers' control over experimental manipulations offers the opportunity to introduce a treatment anew. Modifications and extensions of experimental designs may also offer evidence on operative mechanisms and allow informed discussion of whether those mechanisms are likely to operate in distinct contexts.[25] Replication and carefully scaled-up extensions of experimental designs are thus thought to provide the most reliable route to cumulative learning, and can ultimately inform questions about the conditions under which specific interventions are more likely to work, more likely to be cost effective, or more likely to have unanticipated or unintended consequences.

One can find examples of cumulative learning from replication of experimental research in some domains. These can be categorized into two classes of integrated studies: (a) the same intervention has being implemented in multiple *places*, and (b) the same intervention has been implemented in different ways, for the direct purpose of comparison across different variants of the same class of intervention. In an interesting example of the first class of studies, which is related to the approach we develop in this book, Abhijit Banerjee and colleagues presented findings from six RCTs that tested the effect of a similar program to support the very poor in obtaining access to assets, life skills coaching, savings accounts and health information services. This basic program was adapted to the wide variety of geographic and institutional contexts and was implemented with multiple implementing partners.[26] In other contexts, researchers have taken advantage of a set of high-quality randomized controlled trials that emerged around simultaneously on a

[24] Banerjee and Duflo (2009)

[25] See Clemens (2017) for a discussion of different forms of replication, verification, and robustness tests.

[26] Banerjee et al. (2015).

similar topic to synthesize evidence.[27] An interesting illustration of the second class of integrated studies comes from Tessa Bold and colleagues, who replicated experiments on the educational impact of hiring additional teachers in Kenya but varied whether the hiring was done by NGOs (as in previous research) or through a government ministry.[28] Their results suggested that hiring through a traditional government ministry did not have the same positive impact on educational attainment as hiring through an NGO. In other words, part of the apparent effect of hiring teachers evident in previous research was due to the fact that this was done through a novel, parallel non-governmental structure. In the United States, the large body of experimental research on voter mobilization by Alan Gerber and Donald Green suggests a similar program of replication with controlled variation in design that allows research to build on previous experimental findings.[29] Other researchers and organizations have sought to advance the role of systematic meta-analysis; see, for example, recent work by AidGrade.[30] Such examples illustrate how both experimental replication and meta-analysis can provide a valuable tool in the effort to promote generalizable knowledge.

Yet such replication in fact appears quite rare—and rarely offers these benefits, for several reasons. "Unplanned" replication, to the extent it arises, tends to involve new interventions, distinct outcome measures, and other differences across studies that make it difficult to learn from the comparison of results across studies—the problem we referred to as study heterogeneity. And planned external replication, at least of field-based stud-

[27] See e.g. Banerjee, Karlan and Zinman (2015).

[28] Bold et al. (2016).

[29] Green and Gerber (2015). See also the very interesting meta-analysis and planned replication of campaign mobilization studies by Kalla and Broockman (2017).

[30] http://www.aidgrade.org.

ies appears almost as rare in experimental as in observational research.[31] The reasons are likely many, but one key factor may again be strong career incentives that reward innovation, and weak career incentives to replicate. In addition, as we noted, the rise of experiments has increased the deep involvement of researchers with the intricacies of study sites and the requirements for design implementation in a particular context, which while encouraged, can further weaken comparability across studies. Moreover, researchers tend to emphasize theoretical innovations—to differentiate themselves from others—such that different papers ostensibly on the same topic focus on distinct aspects of the same problem; and different experiments may not only study distinct interventions and theoretical claims but also use different measures and estimation strategies. Indeed, with RCTs (as for observational studies), incentives to innovate extend also to measurement strategies, with professional rewards for novelty. Perhaps for these reasons, as Gerber and Green (2012, 347) put it, "any two experiments differ along an unmanageably large number of dimensions." Uncoordinated replication—itself substantially rare—generates serious challenges to drawing informative conclusions from the aggregation of findings.

The epistemic challenge is therefore how to coordinate RCT research in a manner that doesn't sacrifice internal validity but does meaningfully enhance our understanding of general effectiveness, and of the role played by context, as well as how variations in the implementation of interventions may condition their impacts.

2.2 The "Metaketa" Approach

The barriers to cumulative learning described in the previous section are substantial. The complexity of the social and political world; the organization of knowledge production

[31] External replication of laboratory studies is easier because the lab, by definition, is devoid of context. See for example, Henrich et al. (2004).

in the professional social sciences; and the recent revitalization of RCTs as a mode of strengthening the internal validity of studies collectively pose substantial obstacles.

This book describes our efforts to address these challenges. In particular, we report results of a novel initiative of the Evidence in Governance and Politics (EGAP) research network. The inaugural project of the "Metaketa Initiative" that we present integrates seven coordinated studies led by thirty PIs from nineteen universities.[32] Our motivation stems from the recognition, discussed in the previous section, that individual researchers working independently do not necessarily generate the optimal set of studies for knowledge accumulation: problems of study scarcity and heterogenetiy, selective reporting, and private data can undermine the accumulation of knowledge. Our goal is therefore to generate cumulative evidence by funding field experimental research across disparate contexts, working with independent project teams to increase coordination among studies that share common research questions and hypotheses. We seek to increase the number and comparability (within constraints discussed below) of studies in a given topic area in order to support cumulative learning from a set of studies as a whole. The data generated by such an effort may be more feasibly integrated in an overall meta-analysis than can the output of individual, uncoordinated studies. Variation in findings both across studies and due to planned experimental variation within studies can, potentially, also contribute to addressing several of the specific challenges to external validity in experimental research, including issues of excludability, mechanisms, and scale. A major portion of this effort involves the construction of research vehicles that incentivize replication as well as innovation. Our wager is thus that greater coordination among experimental researchers can contribute to counteracting several of the difficulties described in the previous section.

[32] It was backed by over $2 million in funding; see our discussion of the funding process in the preface and below.

Our objective in writing this book is partially substantive, in that we lay out answers to questions about information and accountability gleaned from the inaugural Metaketa. But it is also methodological, in that we lay out the rationale for and structure of the initiative; attempt to validate its usefulness as a tool for cumulative learning; and suggest guidance and lessons learned for future initiatives of coordinated studies.

In the rest of this chapter, we describe the Metaketa grant-making model and discuss how this approach may in general help overcome some of the challenges to cumulative learning described above.[33] In the remainder of the book, we then assess and demonstrate the model's utility for addressing important and practically relevant questions surrounding the relationship between informational interventions and political accountability. Beyond that important substantive focus, however, we view this inaugural project of the Metaketa Initiative as a replicable model. Indeed, the research reported in this book has motivated three additional EGAP-sponsored Metaketas focused on (1) taxation and accountability, (2) natural resource governance, and (3) community policing, and we hope these will generate additional coordinated experimental research projects that build on this model.[34]

The Metaketa Initiative is based on a number of core pillars, designed to overcome, to the extent possible, challenges both to the reliability of individual studies and, especially, to the credibility of the overall inferences that can be drawn from a set of related studies. We summarize these challenges and the pillars of our approach in Table 2.1. In the next section, we further discuss the rationale for these elements of our approach as well as the steps involved in implementing them.[35]

[33] As noted in the preface and Chapter 1, "Metaketa" is a Basque word meaning "accumulation."

[34] As noted in the preface, Metaketa I was funded by an anonymous donor; the subsequent Metaketas are funded primarily by the United Kingdom's Department For International Development (DfID). Metaketa I was administered by EGAP in conjunction with the Center on the Politics of Development at the University of California, Berkeley.

[35] See also our pre-analysis plan for the meta-analysis, as well as each study's project-specific pre-analysis

Table 2.1: The Metaketa Initiative: Extant Challenges and Pillars

Extant challenges	**Pillars of the Metaketa Initiative**
1. Confounding in observational research	1. Randomized controlled trials
2. Limited external validity of single RCTs	2. Multiple studies in diverse contexts
3. Heterogeneous, scattered findings	3. Meta-analysis with overall finding
4. Diversity of interventions	4. "Common arm" intervention
5. Non-comparable measures, impeding aggregation	5. Harmonized measurement of inputs, outcomes, and controls
6. Researcher incentives for innovation over replication	6. "Alternative arm" intervention
7. Private data	7. Open data and replication code
8. Errors in data or code	8. Third-party data analysis
9. Fishing (data mining, specification searching, multiple hypotheses)	9. Pre-analysis plans with limited number of specified hypotheses
10. Publication bias	10. Publication of all registered analyses

First, all the Metaketa studies employ randomized interventions to identify causal effects (point 1 in Table 2.1). We thus seek a strong basis for causal inference within each study. This in turn provides the foundation for valid inferences about overall effects, when aggregating across studies.

Next, and critically, we seek to consolidate evidence on major questions of scholarly and policy relevance—with an emphasis on cumulative learning, rather than primarily on innovation (points 2-6 in Table 2.1). In doing so, we seek to address several related barriers to the accumulation of knowledge in the social sciences, especially the problems of study scarcity and study heterogeneity discussed previously.

Thus, because any single RCT may have limited external validity, we fund multiple studies on a single topic across diverse contexts (point 2 in Table 2.1). In place of heterogeneous, scattered results, we also aim to produce a meta-analysis, resulting in an overall finding produced from the aggregation of these multiple studies (point 3 in Table 2.1). And to address the diversity of interventions and the proliferation of non-comparable mea-

plan (see Appendix).

sures in a given area, which can hinder aggregation, research teams strive to coordinate on conceptually similar interventions (point 4 in Table 2.1) and commit to measuring the same variables, including key outcome variables, in a similar way (point 5 in Table 2.1). To be sure, what is meant by "similar" is an important and sometimes difficult question (and one we address in detail for Metaketa I in Chapter 3). The core principle, however, is that to the extent possible, differences in findings should be attributable primarily to contextual factors—and not to differences in research design or measurement. Overall, notwithstanding some differences that arise naturally from working in different sites, close coordination of interventions and outcomes significantly increases the plausibility and tractability of meta-analysis. At the same time, basic similarities in the interventions imply that differences in treatments across studies can also be considered, and distinct treatment effects in different contexts can be pre-registered as hypotheses in light of those differences.

Our emphasis on coordination, replication, and cumulative learning raises an important challenge, however: researcher incentives for innovation over replication. We recognize the importance of innovation for the growth of knowledge. Breakthroughs are rightly prized. Yet, as we outlined above, researchers often have weak incentives to verify previous findings with new studies. This may lead to the privileging of "being first" over "being right." A major question is therefore how to address this issue—recognizing the reality that researchers prize innovation in part because individual studies that are deemed innovative are easier to publish. Our approach, as detailed in the next section, is to coordinate research on a "common" intervention arm among all studies included in the Metaketa but also build in planned diversity across studies, through inclusion in each study of at least one "alternative treatment arm" (point 6 in Table 2.1). In this way, research teams generate comparable results that can be integrated through meta-analysis of common interventions—while also allowing for novel individual findings through study-specific interventions.

Additionally, we seek to address several challenges related to selective reporting, as well as private data. Both individual studies and our meta-analysis seek to take advantage of best practices of analytic transparency, including (i) open data and materials (point 7 of Table 2.1); (ii) third-party replication of analyses prior to publication (point 8 of Table 2.1); and (iii) pre-registration of designs and analysis plans (point 9 of 2.1). Funding for Metaketa I was conditional on researchers' agreement to abide by these and other procedures included in EGAP's statement on research transparency.[36] These practices are designed to limit threats to the validity of individual studies—including fishing expeditions or unintentional errors in data analysis—but also the validity of the meta-analysis. Thus, the analysis plans for aggregating results are also pre-registered: we seek to bring public and transparent methods to the accumulation of results from the separate field experiments that comprise our project.

Finally, in the Metaketa model, researchers commit to be part of an integrated publication of the results of all of the studies, as a way to avoid publication bias (point 10 in Table 2.1). Thus, regardless of findings, and particularly regardless of whether the estimated effects of individual studies are statistically different from zero, the results of each study would be published in a single prominent outlet.[37] Joint publication of the results—particularly if negotiated with a publisher in advance of analyzing study data, as is the case in this volume—can involve a form of "results-blind" review in which publication decisions are driven by the quality of the research questions, theory, and research

[36] See http://egap.org/resources/egap-statement-of-principles/.

[37] All project teams are writing, in addition, separate individual articles, often based on the "alternative" rather than "common" arms of their experiments; see further discussion below.

designs, but not the statistical significance of the findings.[38] An integrated publication limits publication bias, because null results readily appear. Even the outcome of failed interventions—as in one of the studies in this volume (Chapter 10)—can appear in the final write-up, rather than disappearing in the file drawer of unrealized projects.[39] As we detail later, with our approach it is not possible to sweep study-level attrition under the rug; and knowing about missingness of outcome data within and across studies is informative and highly useful for the broader social-scientific inquiry. This book constitutes the integrated publication for Metaketa I.

A different core set of principles, discussed in the next section, relates to ethics. Given that Metaketa interventions may focus on sensitive and important areas of governance (for example, democratic elections), ethical concerns—such as "do no harm" principles—constitute a central focus. We discuss ethical concerns in Chapter 3 and elsewhere in the book.

In sum, all studies in this joint initiative aimed to adhere to collective methodological principles, including pre-registration and third-party data analysis. By coordinating on interventions and outcomes and adhering to important principles of transparency and reproducibility, the initiative aims to maximize comparability and the accumulation of knowledge across different studies. Research teams in the Metaketa Initiative work on parallel, coordinated research projects. They collaborate on theory, intervention design, and on both measurement and estimation strategies in order to allow for informed comparisons across study contexts. This coordination of design, interventions, and outcome measures across studies also makes our data "meta-analysis ready" to a much greater extent than would be a standard set of disparate experimental studies. And pre-specification

[38] On results-blind review, see Findley et al. (2016); Dunning (2016).

[39] Karlan and Appel (2016).

of the meta-analysis plan limits the scope for data mining at the aggregation stage. In addition, Metaketa I sought to facilitate an exchange that makes for a much more open and collaborative model of science than is sometime practiced in social, political, or economic research, for instance, through the holding multiple meetings with all of the project teams at multiple research stages—for instance, to workshop research designs, report progress and present final results. Multiple interactions between teams working in parallel also helps handling multiple logistical and professional issues, including those around modes of publication of results. At the same time, teams have some incentive to both be innovative and to "check over each other's shoulders"—enabling a degree of informed scrutiny and constructive criticism that is helpful for scientific progress. Finally, researchers also observe common principles for ethical research in the area of research on governance more generally, and information and electoral accountability in the case of Metaketa I.

2.3 Making Metaketas Work

Operationalizing these principles was perhaps the major challenge of our work. The Metaketa approach is, to our knowledge, unique in the social sciences in its effort to take a large number of independent research teams and forge a collaborative approach that would result in tightly coordinated field experiments in so many countries, with key design and analysis procedures agreed in advance of data collection.[40] Given that, at our first meeting, we did not have an ex-ante consensus on the details of how this coordination was to be achieved, we had to engage in a careful discussion on a set of potentially contentious issues. In multiple meetings and workshops held between 2014 and 2016, the seven research teams and the thirty co-authors of the chapters in this book

[40] See also the Foundations of Human Sociality project that coordinated researchers across 15 sites, all undertaking similar behavioral (laboratory) experiments; Henrich et al. (2004).

therefore collaborated on theory, design, measurement and estimation strategies in an effort promote cumulative of learning across studies.[41]

It is therefore worth describing several key aspects of this operationalization—i.e., "making the Metaketa work." The goal of the Metaketa Initiative is to facilitate research structures that mitigate the threats to cumulative learning discussed above—in particular, practices and strategies that permit movement away from the status quo and that increase our ability to cumulate reliable knowledge from multiple studies on a single topic. Achieving these goals required new structures as well as practical decisions, and here we describe several of the specific opportunities and challenges we encountered in developing the Metaketa initiative, the solutions we found, and the advantages we see for this new method for cumulative learning. We emphasize, however, that our approach has some important limitations; and there are trade-offs to consider with this form of research. Certainly, learning in the social sciences benefits from many approaches, and Metaketas are only one among these. In our concluding Chapter 13, we discuss "lessons learned" and consider the moments at which and the questions for which a Metaketa may be particularly valuable.

Table 2.2 describes the implementation steps for Metaketa I. In many ways, this looks like any other grant-making initiative: a steering committee issues a call for expressions of interest (EOIs) and then a request for proposals (RFP); awards are made, and preliminary approvals from university and governmental partners are obtained; baseline data are collected, interventions are fielded, and endline data are gathered; and research results are written up, published, and disseminated to key stakeholders.

Yet, several aspects of the structure of implementation are particular to this initiative,

[41] We held such workshops in Princeton (October 19, 2014), Cambridge/MIT (December 7, 2014), San Francisco (September 3-4, 2015), and Berkeley (December 8-9, 2016).

and, in particular, to the challenges we faced in developing coordinated research on a single topic with independent research teams, in the context of an effort to privilege especially the role of replication and the pillars of the Metaketa approach we described above. In the rest of this section, we describe the ways in which sought to resolve a tension between innovation and replication, with a focus on consolidation of evidence; bolster research transparency; respect key ethical principles; achieve coordination on interventions and outcome measures across projects; and combine the results of the studies in a formal meta-analysis.

Table 2.2: Metaketa I: Steps of Implementation

Preliminaries	Implementation (preparatory stage)	Implementation (study stage)	Post-implementation stage
1. Establish steering committee 2. Raise grant 3. Decide broad focus area 4. Solicit Expressions of Interest (EOIs) 5. Identify research question 6. Request for Proposals (RFPs) and R&Rs 7. Final award decisions	1. Coordination meetings 2. Write and register study-specific PAPs 3. Seek IRB approvals and government permissions	1. Write and register meta-preanalysis plan (MPAP) 2. Run baseline and interventions 3. Endline data	1. Analyze and publish meta-analysis 2. Publish individual studies 3. Disseminate findings to key stakeholders

2.3.1 Consolidation of Evidence: Innovation vs. Replication

This inaugural Metaketa involved construction of a selection committee—comprised of the five co-authors of this chapter and chaired by Dunning—who would work separately from authors of the individual studies and who would take responsibility for several aspects of the overall research strategy.[42]

Our first objective was to identify research questions that (1) fit within a general predefined substantive area; (2) mattered to both researchers and policymakers; and (3) included interventions and outcomes that could feasibly be harmonized across multiple studies. Another central concern was to identify areas in which sufficient numbers of researchers were working, or were interested in working, that a group of similar studies could be pursued. We thus pursued a two-stage proposal process, first distributing a call for Expressions of Interest (EOI) to allow researchers to indicate their potential participation and their ideas for research projects at a relatively low cost in time and effort. For Metaketa I, our EOI call sought especially to identify potential projects in the areas of both community-based monitoring and informational interventions for political accountability. The latter emerged as the most promising focus, not only because of its substantive interest and importance but also because of the density of researchers interested in pursuing work in this area. We then released a full Request for Proposals (RFP) that focused on the informational theme and solicited much more detailed descriptions of the proposed projects.[43] We selected projects for funding from among the detailed proposals we received, assessing the quality of the proposed research including the strength of the research designs for causal inference (a pillar mentioned in point 1 in 2.1); all of the funded studies proposed randomized field experiments. We also considered especially the substantive fit with the Metaketa I project; and funding decisions were contingent on researchers' agreement to adhere to principles of research transparency, coordination across studies, and other pillars of our approach.

[42] The ongoing Metaketas II, III, and IV alter this structure somewhat; for example, they allow committee members (but not the chair) to apply to the Metaketa as researchers after the initial definition of core themes through an Expression of Interest round.

[43] This RFP was open to all scholars; it was not limited to current members of the EGAP network, nor to those who had submitted an EOI.

The availability of substantial funding in support of this project certainly eased the task of attracting world-class scholars to participate in it.[44] Yet, it was still non-trivial to ensure sufficient interest in participation. After all, our theory of the problem of knowledge accumulation focuses centrally on the incentives of academic researchers to plant the flag, due to strong professional orientation towards novelty discussed in section 2.1. We cannot (nor would we necessarily want to) globally alter the fact that novelty and innovation are prized in standard publication processes. Yet we believed that if incentives to do so are improved, enough scholars may be willing to build in additional research time to coordinate across studies, such that their work better contributes to the accumulation and consolidation of evidence.

One major issue, then, was how to promote incentives to engage in replication. This was plausibly further complicated by the fact that most of the researchers on the Metaketa I project teams were tenure-track but untenured researchers or were advanced graduate students at the time the fieldwork was planned and implemented.[45] At the same time, we also sought to recognize the value of innovation and to prioritize the study of comparative effectiveness—that is, learning about what works best to increase political accountability, and where and why it does so.

Our approach to this inherent tension was to develop a structure that fosters coordination and replication but also leaves sufficient leeway for researchers to innovate—as well as to publish independent articles on their project-specific findings. In particular, we called for proposals for research designs with at least two treatment arms:

[44] This is perhaps especially true in political science, because money for field experiments is scarcer than in adjacent disciplines such as development economics or public health.

[45] Indeed, we also had teams composed of graduate students at the time of application who carried out exceptional execution of their projects (see Platas and Raffler, this volume; also Lierl and Holmlund, this volume). Their participation offered substantial benefits both for the energy and ambition of the project and with respect to the development of younger scholars.

- One “common” intervention arm focused on provision of information on the performance of politicians in a way that was as similar as possible across studies (point 4 in Table 2.1); and

- At least one alternate arm that varied across projects, and that allowed for assessment of new hypotheses as well as comparison of the impact of different kinds of treatments (point 6 in Table 2.1).

Studies included in Metaketa I, for example, used their alternative arms to vary inter alia whether information was provided privately or publicly; the density of treated communities within a constituency; the identity of the messenger providing the information; and the presence of alternate messages that may heighten the perceived salience and relevance of the information (see Chapter 3 and Part II).

We believed that this structure would promote replication and comparability—through the first treatment arm—while preserving room for innovation through the second arm. Thus, we would seek to learn about aggregate effects through meta-analysis of results on the common arm, building on the replications of similar interventions across disparate contexts; and we would also seek to learn from variation in effects, both across contexts and through experimental variation that is internal to each study. By differentiating the projects along the second arm, we would maximize the chances for independent publication of the results of each study, while still providing a baseline common treatment arm that would be replicated across studies. We felt this model would facilitate researcher participation in a sustainable way and thus could be extended to other initiatives, including future Metaketas. The structure of the call for proposals therefore, we hoped, helped to reconcile the tension between innovation and replication in a way that would allow for the *consolidation* of evidence on information and accountability.

In the end, to be sure, innovation in research topics and replication of research results remain in some ways in tension with each other. Research funding mechanisms that are open to topic, choose the best projects on a case-by-case basis, and 'let a thousand flowers bloom' may generate very high-quality studies, and may innovate in new areas of investigation. Yet, they are unlikely to provide a dense body of evidence on a single topic that can cumulate in a manner that is obviously externally valid. Pure replication studies, such as the Banerjee et al. (2015) study analyzing BRAC's Targeting the Ultra-Poor program can be an excellent way of establishing the broad, externally valid impacts of a very specific intervention, but do not allow for innovation in research questions or in product design. The Metaketa approach was designed to strike a balance between these two goals, permitting a kind of crowd-sourcing of topics of interest, using multiple treatment arms to allow researchers to innovate on individual projects, but harmonizing theory, interventions, measurement and estimation in a manner designed to foster cross-study comparability.

2.3.2 Research Transparency

In structuring this Metaketa and considering the reporting and publication of its results, per Table 2.1, we also sought to combine several features of study registration, pre-analysis plans, and results-blind review. It is useful to describe the extent to which these practices are in fact likely to reduce publication bias and selective reporting—which we identified in section 2.1 as important barriers to the cumulative learning.

Study registration may refer simply to documenting the existence of a study in advance of its execution.[46] In principle, it allows description of a universe of planned studies—

[46] On the benefits of different forms of registration, see Humphreys, Sanchez de la Sierra and van der Windt (2013) or Monogan III (2013); for a lucid critique and discussion of possible drawbacks, see Laitin (2013).

which provides a denominator against which one can assess the set of completed or published studies. To date, registration has been somewhat ad hoc, with several different organizations providing third-party registration services.[47] Several political science journals now have a policy of encouraging study registration.[48] However, registration is typically voluntary, and the level of detail about the planned study varies greatly.

Pre-analysis plans, by contrast, typically describe the hypotheses and statistical tests that will be conducted once outcome data are gathered, often in greater detail than study registration alone would require—though there is currently no strong standard for their form and content, and empirically they can involve greater or lesser specificity about the number and kind of tests. Empirically, many pre-analysis plans discuss research hypotheses but are quite vague in terms of the precise operationalization of tests. At the other extreme is Humphreys et al.'s (2011) approach of posting complete analysis code with mock data (see http://egap.org/registration/602), which allows analysts to simply run the code once the real outcome data are collected. This arguably represents best practice, since it leaves little guesswork on the part of readers of a pre-analysis plan of exactly what is intended in the analysis.[49] Pre-specification of tests promotes credible adjustment for multiple statistical comparisons and limits the scope for data mining.

Finally, results-blind review—as the name implies—refers to the practice of reviewing a research report blind to the study's findings. Thus, referees evaluate a journal submission on the basis of the interest and importance of the research question, the strength of the

[47] As of September 2017, the EGAP registry has over 600 designs registered since inception in March 2011 (see http://egap.org/design-registration/registered-designs/). The American Economic Association (AEA) (see https://www.socialscienceregistry.org), the Open Science Framework, and other entities also host large registries.

[48] See, e.g., *Political Analysis*, http://www.oxfordjournals.org/our_journals/polana/for_authors/general.html.

[49] An interesting related approach exists for studies with pilots: code can be pre-registered after analysis of pilot data, which are likely to have similar characteristics as data collected during scale up.

theory, and the quality of the empirical design—but not the p-values of the study. Though still quite rare, the practice has been applied in several venues.[50]

These three forms of pre-specification likely have different capacities to reduce publication bias. Study registration without pre-analysis plans—while it allows *measurement* of the phenomenon, by providing a denominator for the number of studies in a given area—seems unlikely to reduce the bias. Indeed, whatever the true source of publication bias, the mere fact of having announced the existence of a study prior to its execution should not affect its chance of publication, conditional on the p-values. Consistent with this conjecture, Fang, Gordon and Humphreys (2015) find no evidence that the creation in 2005 of mandated study registration in medical journals—which did not, however, require detailed pre-analysis plans or results-blind review—led to a reduction in publication bias.

With pre-analysis plans, the likely impact is subtler and depends on whether the source of publication bias is (1) specification searches or "fishing" on the part of authors; or (2) the preferences of reviewers and editors for statistically significant findings. In principle, pre-specifying the set of tests to be performed limits the scope for ex-post specification searches or "fishing" for statistically-significant effects; and pre-analysis plans may allow for meaningful adjustment for multiple statistical comparisons—without which the interpretation of nominal p-values may be undermined. A complete pre-analysis plan pre-specifies the mode of adjustment for multiple statistical comparisons and thus limits the scope to condition adjustment on realized p-values. However, if journal editors and reviewers simply refuse to publish null estimated effects—perhaps because they find null effects uninformative—pre-specifying the tests will not reduce publication bias.

[50] A recent special issue of the journal *Comparative Political Studies* featured only articles reviewed in this results-blind way, though it allowed both planned research, described prospectively, and completed research that was stripped of discussion of results. This practice may encourage a selection bias in the types of articles submitted, since authors of studies with null effects might be more likely to strip their article of results and apply to such a forum.

By contrast, results-blind review does appear to offer an effective remedy for publication bias: it is impossible for reviewers and editors to condition publication decisions on the p-values if they do not know what the p-values are.

We took several steps to ensure that these dimensions of research transparency were at the heart of Metaketa I. First, the collection of seven studies were registered as part of our research pre-specification. This also therefore records the existence of missing studies or data and consider the impact of such study-level attrition for our inferences. In fact, one of the seven projects planned for this Metaketa did not occur, as described in Chapter 10. In our meta-analysis in Chapter 11, we consider the implications of this missingness for the robustness of our aggregate conclusions. The critical point here is that without such study registration, no record of this study—nor the fact that it went missing—would exist.

Second, the authors of all the individual studies in this volume registered pre-analysis plans in advance of obtaining outcome data.[51] Even more uncommonly, our meta-analysis reported in Chapter 11 also was pre-registered. Critically, the individual-project pre-analysis plans were written after several coordination meetings (on measurement and estimation strategies), and after we collectively wrote the meta-preanalysis plan (MPAP, see Appendix); this helped ensure coordination across studies at the pre-analysis phase (see also related discussion below) and made the ex-post meta-analysis much more feasible. In Chapter 11 and the book's conclusion, we reflect on lessons learned from the execution of the MPAP. Notwithstanding some limitations, we emphasize that the pre-specification of the analysis of pooled data made our studies "meta-analysis ready" to a much greater extent than would be the case with a standard set of disparate experimental studies.

Finally, a core principle that was outlined in our RFP is integrated publication of the

[51] In most cases, designs were also registered before interventions were fielded.

results of all of the studies. Joint publication of the results—particularly if negotiated with a publisher on the basis of the study designs and in advance of analyzing study data—can involve a form of results-blind review in which publication decisions are driven by the quality of the research question, theory, and empirical design, and not the statistical significance of the findings. Thus, integrated publication can be a tool for limiting publication bias, because null results can readily appear. Even the outcome of failed interventions can be informative for the broader themes of inquiry and is thus useful to the broader social-scientific inquiry. As mentioned above, this book—for which we obtained approval for an advance contract on the basis of a book prospectus (i.e., not on the basis of the studies' *p*-values)—is published on the basis of such a results-blind review and constitutes a main integrated publication for the initiative.

We also considered different forms that such integrated publication could take; for example, we discussed the possibility of a special issue with the current editor of one leading political science journal, but that route did not appear promising, in part because results-blind review is not yet established as a norm in social science publication, and in part because the editor was hesitant to consider steps necessary for integrated publication—e.g., an up-or-down decision on the entire package of studies. Moreover, the journal format did not provide sufficient space for important details on the structure of this inaugural Metaketa nor to allow for substantial synthesis across studies. We also have considered a short, synthetic Science-style article with all Metaketa participants as co-authors that would set forth the primary results of our meta-analysis. A book such as this one, however, negotiated with an advance contract, offers several advantages, including the space to pursue the research model and the process of operationalizing it in depth; and the capacity to present in-depth evidence from each study as well as from the overall meta-analysis. We recognize a certain irony that publication of the book may have eased by the fact that this inaugural Metaketa was intellectually novel and hence involved some "planting the flag" on coordinated research. Nonetheless, we believe volumes on the varied substantive topics of future Metaketas are likely to be appealing to top presses as well.

In addition to this integrated publication, and related to the theme of encouraging participation in the Metaketa Initiative, we encouraged individual project teams to pursue publication of articles, reporting especially on the alternate arms of their studies, in leading journals. We hoped that the distinctions between the projects especially on the second intervention arms would ease such problems; but again, the promise of integrated publications was critical for addressing such concerns. Yet, there were substantial issues that arose in trying to make this approach work. Because our collaborative project focused on information provided to voters in advance of elections—with voting and turnout measures constituting the main outcome variables—the timing of projects was tightly linked to the timing of elections across studies. Thus, some projects fielded much earlier than others. For example, the first intervention was fielded in Benin in connection with the March 2015 election, while the last was fielded in Brazil in conjunction with the October 2016 election there. This created some concern among researchers of a "publication advantage" for studies fielded earlier as compared to later studies (reflecting the fact that the research questions and approach would be more novel for the first studies submitting for publication than it would for later studies—again, the returns to novelty provided an important challenge to our collective work). We considered, but ultimately collectively rejected, proposals to force the "early" studies not to submit for publication before all studies were complete; our emphasis was on offering and coordinating integrated publication on the common interventions, but also facilitating autonomy of researchers to pursue independent publication.

Finally on research transparency, we also committed to public data and replication code, in accord with EGAP principles of analytic transparency, in order to give third parties the opportunity to replicate findings within each study before publication. Collaboration together with some degree of competition among project teams encouraged third-party verification of analysis. Moreover, each team's analysis was eventually independently replicated by teams of graduate students at UC Berkeley and Columbia University. These internal replications did reveal various minor errors and discrepancies in data and code, which we could correct before compilation of this volume, thus increasing the reliability of reported results. Reproduction of results using publicly available data raised some questions about the meaning and extent of "third-party replication." For example, our preference wherever possible was to work with raw data, and to record all manipulations to the data in replication files so that the path from data to results would be clear—thereby allowing useful checks on data processing errors and other mistakes. Yet, this is feasible only to an extent. Project teams uniformly employed in-country enumerators and survey firms to gather primary data from respondents, and also merged these with official electoral results and other data. Third-party replication unfortunately does not allow checking the quality of the data gathered in the field, which was the responsibility of project teams and the organizations with whom they collaborated. Our approach is to begin third-party replication with the raw data file(s) obtained from survey firms and other sources; yet the inability of third parties to check the quality of the data collection in the field is one limitation that should be born in mind.

2.3.3 Ethical Principles

We also attempted to codify best practice in terms of the ethical principles to which our studies adhered. We made a shared commitment to the following practices:

First, each of the individual projects was approved by Institutional Review Boards (IRBs) at all of the home institutions of the Principal Investigators working on the respective project. We sought such unanimous approval to avoid incentives for forum shopping (e.g., seeking approval from IRBs thought to be "soft" on appraisal of risks to human subjects). This is a critical baseline requirement. We recognize, however, that IRB approval does not constitute an ethical blank check. Review boards are focused on some kinds of harm—especially, risks to subjects—but they can do so to the neglect of others—for instance, risks to enumerators and other project employees. Moreover, avoiding risk of harm to subjects does not imply that an intervention is necessarily ethical in other ways. Beyond unanimous IRB approval for all projects, we therefore instituted several additional ethical principles, as described in our meta-preanalysis plan.[52]

Second, precisely to overcome the typical IRB blind spot around harm to investigators, we therefore sought to ensure that we would not to put enumerators and other project staff in harm's way. This had substantial implications for several of our projects—especially the planned study in India (Chapter 10). In fact, several of the studies involved situations of potential threat to enumerators, and in each situation project teams had to negotiate the best possible way to minimize risks of harm. We have the sense that such risks sometimes arise in field experimental studies yet are rarely discussed in write-ups; the chapters describing individual study results in this book describe these difficulties.

Third, we sought informed consent from all subjects. In some cases, subjects knew that information they received was provided as part of a research project. This commitment goes beyond the usual informed consent, for instance, that researchers use when eliciting participation in a survey. (We recognized the possibility that soliciting informed consent in this way could generate Hawthorne-type effects—i.e., behaviors that are influenced by

[52] See Appendix.

the simple fact of being studied—and discuss this issue elsewhere).

Fourth, we sought partnerships wherever possible with local civil society actors (or government bodies, which was the case for the Brazil study) who implemented the experimental interventions. As described in more detail in later chapters, these partners were either already implementing informational interventions similar to those we proposed; or the experimental interventions were consistent with their core missions and activities. Thus, in many cases, these partners, rather than researchers, implemented the interventions (though in collaboration with researchers who designed protocols for randomization and developed other research design details). Also consistent with an idea of country ownership and public transparency, we have made core project data publicly available in primary languages, aiding in the effort of communicating the existence and intent of the project to participants as well as others in the study countries.[53] And several of the research teams have conducted public events in their project countries disseminating the results of their studies. We also avoided interventions that would fall afoul of local laws.

Finally, we elaborated research designs to ensure to the maximum extent possible that our studies would not affect *aggregate* election outcomes. Our interventions are all designed to be non-partisan, in that they do not seek to privilege a single party or candidate. Also, researchers sought and received approval from the relevant electoral commission wherever appropriate. We did not require consent from politicians about whom information was provided, even though these may be affected by the interventions; as our pre-analysis plan suggests, "the principle is that any information provided is information that exists in the political system that voters can choose to act upon or not and that this information is provided with consent, in a non-partisan way, without deception, and in cooperation with local groups, where appropriate." Some of our project teams sought and

[53] See http://egap.org/research/metaketa/.

received consent from high-level political actors, however, as a way of facilitating project implementation.[54] We further discuss specific measures we took to operationalize these ethical principles—and some of the challenges that arose—elsewhere in the book.

2.3.4 Coordination of Interventions and Measures

One of the most difficult issues involved our attempt to harmonize intervention and outcome measures across contexts, to the extent feasible. This was the focus of several meetings, at which all participants collectively debated and planned how coordination would actually take place. In principle, some issues were already settled at this point, as they were spelled out in the RFP and were a condition of funding and participation. Our early coordination meetings dealt with several other issues of design, data collection, and analysis, which bear on the meta-analysis in Chapter 11 as well as the individual studies presented in Part II of the book. We describe a few of these now; in Chapter 3, we lay out more fully the specific hypotheses, analysis, measurement and estimation strategies that emerged from our workshops.

Perhaps the most fundamental issue for consideration is what it means to coordinate on—and even to "harmonize"—interventions. In what ways are experimental treatments to be considered similar—and, critically, at what level of generality? Here we appeal centrally to Sartori's idea of a "ladder of abstraction."[55] In Sartori's approach, the generality of concepts is heightened by decreasing the number of properties that an empirical observation must satisfy to be considered an instance of the concept—that is, by reducing the intension or connotation of the concept. This more minimal intension in turn increases the number of objects that can be thus classified, that is, it increases the extension or

[54] See studies in Part II for details.

[55] Sartori (1970); see also Teune and Przeworski (1970) on comparative, contextualized measurement.

denotation of the concept. These ideas bear centrally on the question of similarity of interventions. At the lowest levels of abstraction, it is clear that any two interventions—particularly those taking place in contexts as distinct as six different countries in Latin America, Africa, and South Asia, as in Metaketa I—must differ on an almost infinite number of dimensions. Thus, as the attributes needed to define a "common" intervention multiply, the number of cases to which the underlying idea can feasibly apply must clearly diminish—illustrating Sartori's tradeoff between intension and extension of a concept. Yet, by focusing on just a few core attributes of commonality and thereby climbing the ladder of abstraction, the generality and extension of the concept can increase. As we discuss in Chapter 3, a central focus in our common study of informational interventions is the distinction between good and bad news, which we operationalize by studying the relationship between voters' prior beliefs and the information that is provided to them about candidate performance. A concept such as good and bad news about political candidates is at a middle level of abstraction and can be meaningfully defined across quite disparate contexts, even if the particularities of the interventions differ substantially. If theoretical expectations are defined at this level of abstraction, then empirical aggregation of results can also be feasible and meaningful. Defining a sensible middle level of abstraction was a critical outcome of our effort to coordinate interventions in Metaketa I and seems a critical component of the Metaketa approach.

A second issue on which we attempted to achieve coordination was the unit of intervention and how this mapped to the corresponding primary outcome of voter choice. Several of the co-authors of this chapter felt strongly that polling station level data was the gold standard, in the sense that self-reported voter behavior when individuals had just been told good or bad information about politician behavior is potentially strongly subject to social desirability bias. This in turn could lead to a spurious treatment effect that would be manifest in what voters said they would do, but not in how they actually voted. For projects that inherently take place at the community level, this is not problematic and only required the ability to map the spatial structure of implementation to the corresponding polling stations; however, plausible and meaningful ex-ante hypotheses of non-zero treatment effects require sufficient density of treatment at the cluster level, as was the case with the assignment of polling stations in the Mexico study. For projects such as Uganda 2 that use a delivery channel that is inherently individual (SMS), this was more of a challenge.[56] Even projects such as those in Brazil or Burkina Faso that were unable to tie outcomes to polling station data were encouraged to think in creative ways about how to use individual preference measurement strategies that were as free as possible of social desirability bias, such as their use of blinded "ballots" through which voters recorded their preferences over candidates.

Another issue involved the coordination of outcome measures and covariates. In order to permit meta-analysis on any our hypotheses, and particularly in order to conduct analyses of heterogeneous effects in our pooled study group, we needed to ensure consistency of measurement. Each study had a definition of "good" and "bad" news (regrading incumbent performance) that, while quite distinct across studies, was appropriate to the context and based on objective information about performance (see Chapter 3). Trickier was how to define concepts like "receipt of clientelistic benefits" in a manner that was consistent across contexts. With some types of heterogeneity, such as the role of coethnicity in driving the response to politician information, we simply recognized a priori that these concepts would be relevant in some contexts and not others (specifically, Sub-

[56] The implementation of the randomized saturation design undertaken in the Uganda 2 project was the result of an attempt to achieve coordination in this respect, because this generated cluster-level variation in individual treatment intensity that could then conceivably move outcomes at the polling station level. Note however that the variation in saturation is among sampled respondents, but they tend to be a small proportion of the communities in which they are nested; see Chapter 7, this volume.

Saharan Africa and South Asia, but probably not the Latin American studies). We also pre-specified a set of hypotheses that we explicitly recognized would vary across studies due to the heterogeneity in context and implementation.[57] Study teams shared survey instruments freely; the entire group worked to discuss the instruments used by the first team to implement its study in the field (Benin); and this instrument then served as a kind of template for measurement for some of these concepts in the other projects.

We also debated in these workshop many other issues, including the form of eventual publication of the results of the studies, and we held several meetings at which interim and final results were presented. This provided ample opportunity for another aspect of the Metaketa Initiative discussed in section 2.2, collaboration, as well as some degree of healthy competition. Indeed, these exchanges across numerous researchers before, during, and after the fielding of interventions enabled a degree of informed scrutiny that is helpful for scientific progress, and made for a more "open" mode of science than is often practiced in social, political, or economic research.

2.3.5 Formal Synthesis

In this Metaketa I, we sought to build learning in an area where there is already some evidence base—and even to model interventions on those found in prior research. Teams attempted to build conceptually similar treatments and to measure similar variables, including key outcomes. To the extent possible, we hoped that differences in findings would be attributable primarily to contextual factors—and not to core differences in research design or measurement—and would therefore better contribute to understandings of what works where, why, and when (see Part II). Most importantly, interventions would be conceptually similar in key ways, such that the average effect of informational interventions on

[57] These are hypotheses H12-H16; see MPAP in the Appendix.

political accountability, for the units in our study group, would be a meaningful quantity.

But what do we mean by “the” effect of information, and for which population of units can we characterize such an effect? This was a subject of considerable discussion and even some cordial disagreement among participants in the Metaketa project, and especially members of the selection committee. One important issue was how to conceptualize the study group for the combination of the seven planned studies. Perhaps the simplest approach, and the one that involves the weakest assumptions, is simply to condition our inferences on the set of units and respondents that made their way into the study groups for each of the seven studies. Thus, these units/respondents were not usually drawn from a straightforward probability sample of the population of each country (e.g., because particular regions or states were selected purposively, and/or because some types of respondents were screened out in each study prior to random assignment to treatments). And much more obviously, the six countries included in the Metaketa project are not themselves drawn from the population of countries through some chance procedure but were rather the outcome of the selection process after our RFP was circulated and proposals were submitted. The selection of these countries reflects myriad factors. We certainly hoped for regional diversity but the quality of the projects and research teams was also paramount in driving funding decisions. Thus, it may strain credulity to suggest that these particular six countries, and the effects obtained in each of them, represent some random draw of the possible effects that could be obtained in informational interventions of the sort we consider.

In this view, then, our inferences will be drawn to the particular set of units/respondents in our study: we have a large experiment that pools units in the different contexts and assigns them to treatment conditions in a randomized way. As our pre-analysis plan puts it, “The most straightforward way to combine results across the seven studies pools units into one large study group and estimates treatment effects, as one would do in a large experiment in which treatment assignment is blocked. For this analysis we proceed as if blocking is implemented at the country level."[58] This is similar to many social-science experiments, in which the study group cannot be conceived of as a meaningful random sample from some larger population, and inferences are drawn to the study group.

An alternative approach seeks to conceive of these cases as part of a larger family of possible cases. The goal here is to use each study to contribute to learning about how things work in the family as a whole in order to be in a better position to learn something about each cases from patterns seen in other cases as well as to say something about new cases that have not yet been studied. The ambitions of this approach are clearly much greater but the assumptions needed to justify it are also much stronger, and too strong for many. Most critical is a willingness to assume that the cases can be treated as if they are a draw from a well defined population of cases.[59] For skeptics this kind of analysis might be best thought of as a thought experiment: what kind of general claims could one make if the assumption held? Our pre-analysis plan describes as a supplementary analysis a model for Bayesian hierarchical analysis that provides a way to combine information across the studies in this way. The basic approach assumes a data generating process in which average treatment effects in each case are a draw from a distribution of average treatment effects in a population; the quantity of interest becomes the mean and variance of this distribution; the former provides a best guess for effects in a new case, the latter provides a handle on how much we should expect effects to be similar across cases.[60] In Chapter 11, we further describe our formal synthesis of the experiments in our study.

[58] MPAP; see Appendix.

[59] In principle one can alternatively incorporate beliefs about sampling probabilities of a case into study and assess robustness of conclusions to this.

[60] MPAP, 12–16; see Appendix.

Several important considerations arose in the course of our pre-planning, for example, about whether or not adjust for covariates (we pre-specified both analyses). We further discuss the tradeoffs involved in different choices in Chapter 11 as well as in our conclusion.

We recognize that our focus on aggregating effects has some costs and involves several tradeoffs. Overall, in thinking through the pros and cons of this type of highly coordinated research activity, we might draw an analogy between centrally planned economies and centrally planned research endeavors. This modality empowered us, as the Metaketa coordinators, to direct research effort and money towards specific interventions and specific ways of testing hypotheses. While this tightly integrated approach led to a much clearer cumulation of evidence than a looser approach would have, it did not "let a thousand flowers bloom." A more loosely coordinate research endeavor might be a more effective way of discovering the unexpected, and may allow for a higher average quality of project to be funded than endeavors where only projects cohering to a specific project are funded. Hence it is worth recognizing explicitly that this coordination, while worthwhile, came at a cost. The opportunity cost of this approach is the lost creativity of the more varied ideas that would have come out of our project PIs had they not been required to coordinate so closely, and other innovative and impactful research ideas that were not funded because the topic was not sufficiently close to the one we settled on as the subject of the Metaketa. Moreover, it is clearly the case that not all types of studies lend themselves to coordination, especially when say government agencies are involved in creating and implementing the interventions. We return to these ideas and a discussion of the tradeoffs later in the book, and in Chapter 3, we discuss issues specific to aggregating the effects of informational interventions.

In sum, the challenges of operationalizing our research objectives were considerable. Perhaps the biggest obstacles involved creating a structure that facilitates coordination while also respecting innovation, and creating a plan for publication that would sufficiently reward individual teams for their work while also putting central focus on the coordinated effort. Closer to the substance of the research, there were also substantial decisions to be made around the specific hypotheses, research designs, and measurement strategies. In our several meetings with the full group of Metaketa participants, we considered such issues and also presented designs and workshopped pre-analysis plans. These fora also had substantial ancillary benefits, in terms of fostering a sense of commitment to the shared enterprise and also in pre-committing teams to research hypotheses and tests in a quite public way. Indeed, one outcome of this public pre-commitment to research protocols is that deviations would be observable (and would require explanation to the community). Overall, however, perhaps the biggest benefit–and most important departure from usual practice—is the relatively open and collaborative model for social science that the structure helped create. Our approach certainly has drawbacks and limitations, and involves some tradeoffs vis-a-vis other approaches that we discuss later; yet we feel this open and public style of research focused on creating different modalities of replicability has important benefits. And for us as researchers, it involved a refreshing and important change of pace.

The rest of this book reports the results of this collaboration, and also our effort to validate whether this approach improved cumulative learning. In the next chapter, we turn to our substantive focus—the impact of informational interventions on voter behavior, as well as electoral accountability.

Chapter 11

Meta-Analysis

ANIRVAN CHOWDHURY, THAD DUNNING, GUY GROSSMAN, MACARTAN HUMPHREYS, SUSAN D. HYDE, CRAIG MCINTOSH, AND GARETH NELLIS

Does information shape electoral choices and thereby promote political accountability?

The chapters in Part II of this book provided answers to this core question in particular contexts. The studies individually provide rich insights not only into the impact of interventions that were common to all studies, but also on the effects of alternative interventions that were specific to each study. In addition, several of the studies generated several intriguing hypotheses about heterogeneous effects across respondents with different characteristics.

Yet, what light do these studies collectively shed on the overall impact of informational interventions? And what do they tell us about variation in impact across disparate contexts? Although there are important differences across our studies, they also have critical commonalities; importantly, they all seek to assess common hypotheses about the impact of harmonized informational interventions, using consistent measurements of outcome variables. The coordination involved in constructing these studies substantially addresses problems of study scarcity and study heterogeneity.[1] And it allows us feasibly to pool separate results into a single meta-analysis.

In this chapter, we therefore assess the overall lessons that we can glean from our coordinated studies. This chapter also begins Part III of the book, where we focus on cumulative learning. Our analysis allows us to assess whether, across the set of studies in the initiative, information about politician performance led voters to alter their electoral choices. It also informs a discussion about the conditions under which they did or did not do so, for example, by testing idiosyncratic findings about heterogeneous effects from individual studies on out-of-sample data drawn from the full set of completed studies. Drawing on the merged dataset also allows us to address the robustness of our results—for instance, to assess whether the failure to complete one of the studies (see Chapter 10) could plausibly affect our overall conclusions.

In brief, we find that the overall effect of information across all studies is quite precisely estimated—and not statistically distinguishable from zero. While the results show modest impacts of information on voters' posterior beliefs or political knowledge, informational interventions did not appear to shape their evaluations of candidates—and in particular did not discernibly influence voter behavior. Similarly weak overall findings hold for voter turnout. Moreover, using the merged data and the specifications included in our pre-specified meta-analysis plan (MPAP), none of the country studies individually shows statistically significant effects of information on voters' support for incumbent politicians.[2] We explore several reasons for our findings in this chapter. However, the fact that our results are so consistent across the individual studies limits the possibility that our null effects are due to idiosyncrasies in implementation or study design or limited statistical power. The consistency of our country-specific and the overall results also underscores the value of funding multiple related studies: we are able to reach a more robust overall conclusion than we could have reached from one or two studies

[1] See Chapters 1 and 2 for discussion of these problems.

[2] As we discuss later, pre-specified differences in operationalization and analysis of the common datasets result in minor differences between country-specific analyses in our meta-analysis and several results reported in the chapters of Part II.

alone.

To be sure, our findings do not indicate that voter information campaigns are always ineffective. Evidence from the alternative intervention arms suggests hypotheses about conditions under which information may have more impact. Such findings demand further study—and may suggest the value of future Metaketas focused on assessing such evidence more systematically. Yet overall, our results suggest important limits on the capacity of informational interventions to boost electoral sanctioning and thereby enhance accountability.

In the rest of this chapter, we first describe the pre-specified approach that we use to analyze the pooled data set. We then report our main results and robustness checks; describe the consistency of results across studies; and assess several plausible reasons for our null findings by testing our pre-specified hypotheses as outlined in our MPAP. This chapter could be profitably read in conjunction with Chapter 3, which discusses the common interventions and our measurement of key variables, but it can be read as a standalone chapter as well.

11.1 Primary analysis: Average Effects Across Cases

11.1.1 Hypotheses and Estimation

In previous chapters, we described the core theories of political accountability that motivate our focus on information and electoral behavior. As outlined there, each of the informational interventions in our Metaketa focused on the performance of politicians or their parties. Thus, six studies provided information related to incumbents' legislative performance (Benin), spending irregularities (Brazil, Mexico, and Uganda 2), the quality of public services in their jurisdictions (Burkina Faso), and their quality as candidates (Uganda 1).[3] In their common intervention arms, each of the studies sought to disseminate publicly available performance information that

[3] A planned seventh study on incumbent criminality in India did not take place due to implementation challenges (Chapter 10).

is directly attributable to an incumbent candidate or party; to provide this information privately to individuals within a month prior to an election; and to divulge performance information that is presumed to be relevant to voter welfare. In their second, complementary intervention arms, studies also varied the medium for information provision; the kind of information provided; or the scale of the information provision, for example, by providing information publicly to groups instead of privately to individual voters. See Table 3.1 in Chapter 3 for a summary.

We focus with one exception on analysis of the common arm in this chapter, as registered in our MPAP.[4] Critically, each study was designed to allow measurement of the extent to which voters update their beliefs about the performance of the politicians positively or negatively in light of the information—and to allow measurement of the difference between prior beliefs and provided information. As described in Chapter 3, we expected effects to derive from *new* information rather than *any* information. Most teams gathered information on voter priors at baseline (in both treatment and control groups) with respect to the information that would be provided.[5] Where possible, prior beliefs were gathered on the same scale as the information that was eventually provided to individuals assigned to the treatment groups. This allows us to identify voters who would have received positive or negative information, if assigned to the treatment group. Our empirical strategy therefore takes account of both the content of the information and prior beliefs.

Our core hypotheses for meta-analysis thus concern the impact of positive and negative information (or "good" and "bad" news, see Chapter 3) on vote choice, as well as turnout. We pre-registered two primary hypotheses related to electoral behavior:

- H1a: Positive information increases voter support for politicians (subgroup effect).

[4] See the book appendix.

[5] As discussed in Chapter 3, the Mexico study did not conduct a baseline survey, due to prohibitive costs. The Mexico team instead gathered information on voter perceptions of incumbent malfeasance in the control group at endline, then showed respondents the treatment flyers and estimated the average change from prior to posterior at the randomization block level (see discussion below and Arias et al. Chapter 5). The Burkina Faso (Chapter 8) and the Brazil (Chapter 9) teams measured priors and then administered treatment as part of the same (baseline) survey.

- H1b: Negative information decreases voter support for politicians (subgroup effect).

These hypotheses are straightforward, yet critical: as discussed in Chapter 3, they are necessary components of many models of electoral accountability.

We also registered secondary hypotheses related to electoral participation:

- H2a: Good news increases voter turnout.
- H2b: Bad news decreases voter turnout.

While this hypotheses are directional, our interest is in estimating the relation, whether it is positive, negative, or context dependent. In sections 11.1.2 and 11.1.3, we describe further hypotheses about the impact of our informational interventions on intermediate outcomes, such as perceptions of candidate integrity and effort; the possibility that politicians would mount campaigns in response to negative information; and the conditional effects of information, depending for example on co-ethnic and partisan ties between citizens and politicians.

As described in Chapter 3, the most straightforward way to test our primary and secondary hypotheses across studies is to divide subjects into groups based on whether they would receive "good" or "bad" news if exposed to the treatment. These are sub-group effects, because the groups are defined according to subjects' prior beliefs as well as the provided information. We can then pool the respondents in these good- and bad-news groups across the different studies in Metaketa I. Thus, it is as if we estimate effects in a large, pooled experiment, with treatment assignment blocked by country.[6] This approach involves weak assumptions, compared to alternatives. For example, we do not conceive of the study group of subjects as a random sample from a larger population: the study sites (countries and locations within countries) are not random draws from a well-defined population of possible sites. To be sure, pooling does imply that we can treat interventions and outcome measures as sufficiently comparable that an overall average treatment effect—say, the effect on vote choice of exposure to "good news"—is

[6] That is, units were grouped into "blocks" and random assignment was conducted separately within each block. In our analysis, blocking is in the first instance at the country level; but there is also blocking within countries, as a strategy for reducing the variance of treatment effect estimators.

meaningful. Creating such comparability is one goal of the Metaketa Initiative, and the harmonization of measures of good and bad news across contexts makes an important contribution in that regard.

We focus in the first instance on a particular estimand: the average of the country-specific effects, i.e., the average of the average treatment effects from each study. This approach permits us to assess a single causal parameter across studies but also allows for natural investigation of heterogeneous effects across study sites. In each study, the average treatment effect is the average potential outcome under treatment (exposure to the common information arm), minus the average potential outcome under control. This parameter is unobservable, because we cannot assign all respondents to both the treatment and control conditions, but random assignment gives us two random samples—the treatment and the control groups, respectively—which we use to estimate the average potential outcomes under both conditions. We can then estimate average treatment effects by comparing the mean difference between the treatment and control groups, sometimes weighting the countries as described below. In one set of analyses, this comparison is unadjusted by covariates, other than fixed effects for treatment assignment blocks. We also estimate the simple average treatment effect across the pooled study group of all respondents, rather than the average of the average effects in each country.

In addition, as also pre-specified in our MPAP, we present covariate-adjusted analysis including a full set of treatment-covariate interactions.[7] While our estimands are all defined under the Neyman potential outcomes model, the average treatment effect can be conveniently estimated by fitting two regressions, one for the good news and one for the bad news group:[8]

$$E(Y_{ij}|i \in L^{+}) = \beta_0 + \beta_1 N_{ij}^{+} + \beta_2 T_i + \beta_3 T_i N_{ij}^{+} + \sum_{j=1}^{k} (\nu_k Z_i^k + \psi_k Z_i^k T_i) \tag{11.1}$$

[7] See discussion in Lin (2013).

[8] In the Neyman model, potential outcomes under treatment or control are fixed for each respondent but are free to vary across respondents; see Splawa-Neyman, Dabrowska and Speed (1990), Rubin (1978), and Holland (1986). The only random element in this model is the stochastic assignment to treatment or control.

and

$$E(Y_{ij}|i \in L^-) = \gamma_0 + \gamma_1 N_{ij}^- + \gamma_2 T_i + \gamma_3 T_i N_{ij}^- + \sum_{j=1}^{k}(\nu_k Z_i^k + \psi_k Z_i^k T_i). \tag{11.2}$$

Here, T_i is the treatment assignment variable, and N_{ij}^+ and N_{ij}^- are the gaps between priors and information in each group, standardized to have zero mean and unit variance. Thus, $N_{ij}^+ \equiv Q_j - P_{ij}$, given that $Q_j - P_{ij} > 0$, where Q_j is the provided information about politician j and P_{ij} is voter i's prior belief about politician j, on the dimension about which information is provided. A voter i is in the "good news" group ($i \in L^+$) when performance exceeds her priors, or when performance information confirms positive priors: that is, $Q_j - P_{ij} > 0$, or $Q_j = P_{ij}$ and Q_j is greater than the median performance in the relevant locality. Otherwise, she is in the bad news group ($i \in L^-$). Furthermore, $Z_1, Z_2, ..., Z_k$ are prespecified covariates, also standardized with zero mean. Given the mean-centering of all variables, β_2 denotes the average treatment effect of information for all voters receiving good news; and γ_2 is the average treatment effect of information for all voters receiving bad news. When Y_{ij} measures support for the candidate or party about whom information is provided, then according to our primary hypotheses we expect $\beta_2 > 0$ and $\gamma_2 < 0$. We estimate equations (11.1) and (11.2) by OLS, adding fixed effects for constituencies or for the blocks within which random assignment occurred (when appropriate). This is akin to estimating a linear probability model, which we do for ease of interpretation of the coefficient estimates.[9]

To test a secondary hypothesis that information effects are stronger when the gap between voters' prior beliefs about candidates and the information provided is larger, we combine data from the good and bad news groups and estimate more simply:

$$E(Y_{ij}) = \delta_0 + \delta_1(Q_j - P_{ij}) + \delta_2 T_i + \delta_3 T_i (Q_j - P_{ij}). \tag{11.3}$$

[9] Substantive results do not change if we instead use probit or logit models.

In our MPAP, we expected $\delta_3 > 0$ but noted important caveats about this analysis. For example, our measures of $Q_j - P_{ij}$ are largely ordinal not interval; as we noted in our MPAP, estimating a linear marginal effect of the gap may not be meaningful if the marginal effect is not in fact linear. Again, we do not manipulate priors in our experiments, and we lack an identification strategy that would allow us to make strong causal claims about the effects of such a gap.

As pre-specified in our meta-pre-analysis plan (MPAP), we present both unadjusted and covariate-adjusted results. There are tradeoffs involved in covariate adjustment. Precision gains using a full set of covariate-by-treatment interactions with mean-deviated covariates may be substantial, and the small-sample bias from regression adjustment diminishes rapidly.[10] Moreover, in contrast to many settings, here full and transparent pre-specification of the covariates used for adjustment removes, in principle, the possibility of data mining and specification searches. However, implementing covariate adjustment across projects is not trivial, in part because covariates must be gathered and measured in similar ways across studies. For example, we pre-specified a list of fourteen covariates in our MPAP but in the end project teams could only measure ten of these symmetrically across all studies.[11] Finally, as pre-specified, we impute missing values of covariates using the average value of the covariate in a specific block.[12] We present results without imputing and with imputing missing values.[13] Unadjusted results have the advantage

[10] Freedman 2008a, b; Lin 2013.

[11] The covariates used in the results in this chapter include measures of N_{ij}, age (M14), years of education (M17), wealth (M18), whether the respondent voted in last election (M20), whether voted for incumbent in last election (M21), exposure to clientelism (M22), perception of the credibility of the information source (M24), baseline belief in secret ballot (M26), and whether the respondent perceived the election as free and fair (M27). Here, we give in parentheses the measure numbers used in the MPAP; see Appendix A.

[12] Thus, following the MPAP, we impute the average values of the covariates in the control group in lowest randomization block for which data are available missing data. Note that there are still some missing values after imputation, reflecting observations for which no data on a particular measure is available in the control-group block. In addition, we only include those covariates for which we have some covariate data from every study.

[13] Note also that our MPAP specified that we would report study-by-study F statistics for the hypothesis that all covariates are orthogonal to treatment, using the full set of baseline covariates described in that document. See individual studies in Part II for balance tests.

of simplicity, and it is easiest to hew closely to the pre-specified analysis when covariates are not included. Moreover, in an experiment as large as ours, the precision gains from covariate adjustment are often minimal; and unadjusted and covariate-adjusted estimated effects and standard errors differ little.

In other respects, the analysis in this chapter also closely follows our MPAP. We note three deviations, however. First, we specified that we would cluster standard errors on politicians (j) but this was clearly a mistake in our pre-specification; random assignment occurs within politicians, and so any clustering should reflect that design. Where treatment assignment is clustered, our analysis reflects that (i.e., standard errors are clustered at the level of assignment).[14] In this chapter, we use outcome data at the individual level.[15]

Second, while our pre-analysis plan is not everywhere clear on this point, we intended to conduct hypothesis tests by randomization inference (RI), and we present RI-based p-values in the analysis that follows. Here it is critical that the RI procedure follows the randomization schemes used in each of the studies, including any clustering or blocking of randomization. Thus, we simulate the permutation distribution of estimators such as $\widehat{\beta}$ or $\widehat{\gamma}$ under the strict null hypothesis of no unit-level effects, where the permutations are drawn from the possible realizations of random assignment given the block-randomized or cluster-randomized design in each study.[16]

Finally, we did not clearly specify a weighting scheme in our pre-analysis plan. Yet, as discussed above, we are interested in two estimands: the average of the average treatment effects in each of our studies; and secondarily, in the average effect across all respondents in the studies. To estimate the first parameter, we weight by the inverse of the ratio of the country

[14] For instance, the unit of randomization in Adida et al.'s study of Benin (Chapter 4) was the rural village or equivalent urban quarter; in Arias et al.'s study of Mexico (Chapter 5), it was the precinct.

[15] Only some studies collected aggregate data (e.g. on official electoral results), and we do not pool those analyses here. Some of the individual studies presented in Part II (e.g. the Arias et al. study in Mexico; see Chapter 5) do present analysis of aggregate data.

[16] See our replication code for details.

study group size to the pooled study group size, so that smaller studies are upweighted and larger studies are downweighted. This approach equalizes the contribution of each study to the overall estimate and prevents larger studies from being arbitrarily upweighted in our estimation of the average study-level effect of information. To estimate the second estimand, we instead pool the data without weighting; this approach relies more heavily for the overall estimate on higher-powered studies. We present results in the next section with and without weighting.[17]

11.1.2 The Effect of Information on Electoral Support

How, then, does performance information affect electoral performance?

Figure 11.1 shows the average effects of the informational treatments. Here, we condition on priors by dividing the study group in the good news and bad news samples. And following our discussion in the previous section, we present results from three different specifications:

1. Unadjusted treatment effect estimators (a difference of means), giving equal weight to the six studies;[18]

2. The full specifications in equations (11.1) and (11.2), including N_{ij} and a full set of covariate-treatment interactions, also giving studies equal weight; and

3. An unweighted version of equations (11.1) and (11.2), where the influence of the studies reflects their sample sizes.[19]

In brief, the effect of the informational treatments on vote choice is quite precisely estimated—and null. Thus, across the 15,246 respondents in the study group for the unadjusted analysis

[17] Note that in principle, we could also use inverse probability weighting, in particular, the product of country and inverse-probability weights, to account for unequal probabilities of treatment assignment across countries. However, once we subset to the control group and common intervention arm, we have common treatment assignment propensities across our studies.

[18] These regressions include fixed effects for treatment assignment blocks, as do other regressions.

[19] Note that for the covariate-adjusted analyses, we impute missing covariate data using our pre-specified procedure. Results are substantively identical without imputation; see the Online Appendix.

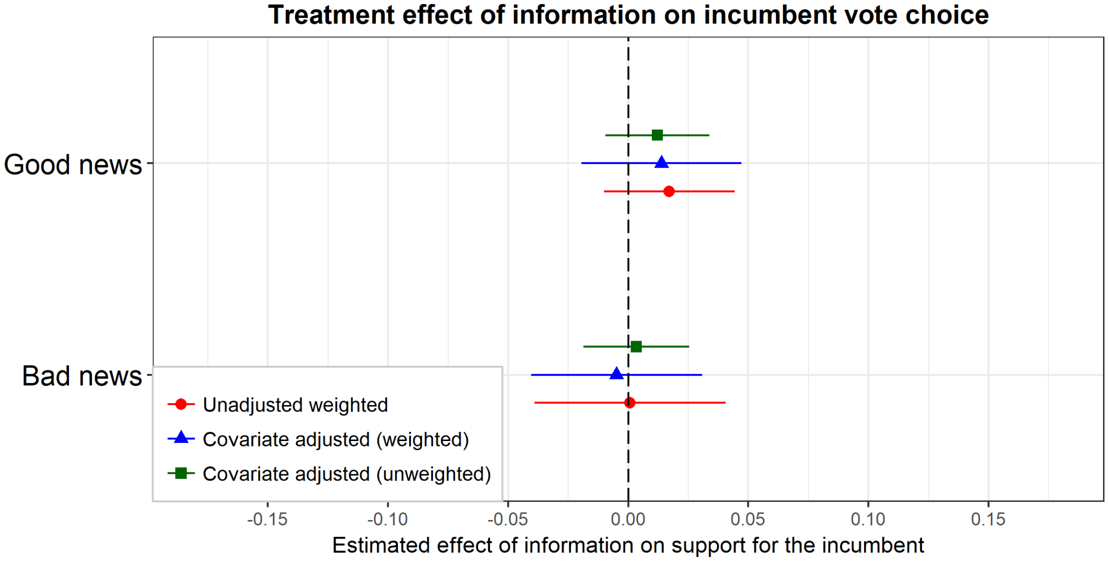

Figure 11.1: Estimated change in the proportion of voters who support an incumbent after receiving good news (top row) or bad news (bottow row) about the politician, compared to receiving no information. The top two estimates in each row are the pre-specified covariate adjusted analyses, with a full set of covariate-treatment interactions (country-weighted analysis in green squares, unweighted in blue triangles); the bottom row shows unadjusted, weighted estimates (red circles). Horizontal lines show 95% confidence intervals for the estimated change. In all cases, the differences are close to zero and statistically insignificant.

(7,417 in the good news group and 7,829 in the bad news group), we see null estimated effects of information in both the good and bad news cases.[20] The results are stable across estimation strategies. The first two columns of Table 11.1 also presents coefficient estimates, standard errors, and randomization-inference based p-values for the observed test statistic, that is, the estimated coefficient on the treatment variable, for the covariate-adjusted analysis. We discuss the robustness of these results later, but approaches other than the three in the figure—e.g., unweighted, unadjusted analysis—lead consistently to the same substantive conclusions. In sum,

[20] The number of observations rises to 11,644 in the good news group and 12,151 in the bad news group, when we include votes for LCV councilors as well as chairs in the Uganda 2 study; see Buntaine et al., Chapter 8. In that analysis, each respondent in the Uganda 2 study enters twice, and we cluster the standard errors at the individual level.

we find no evidence of an effect of information, in quite precisely estimated regressions using data from six of our seven planned studies.

We also find similar, precisely estimated null results for our main secondary outcome, voter turnout (Figure 11.2). Point estimates are almost exactly zero for each of our three analytic strategies.[21] The third and fourth columns of Table 11.1 give further details.

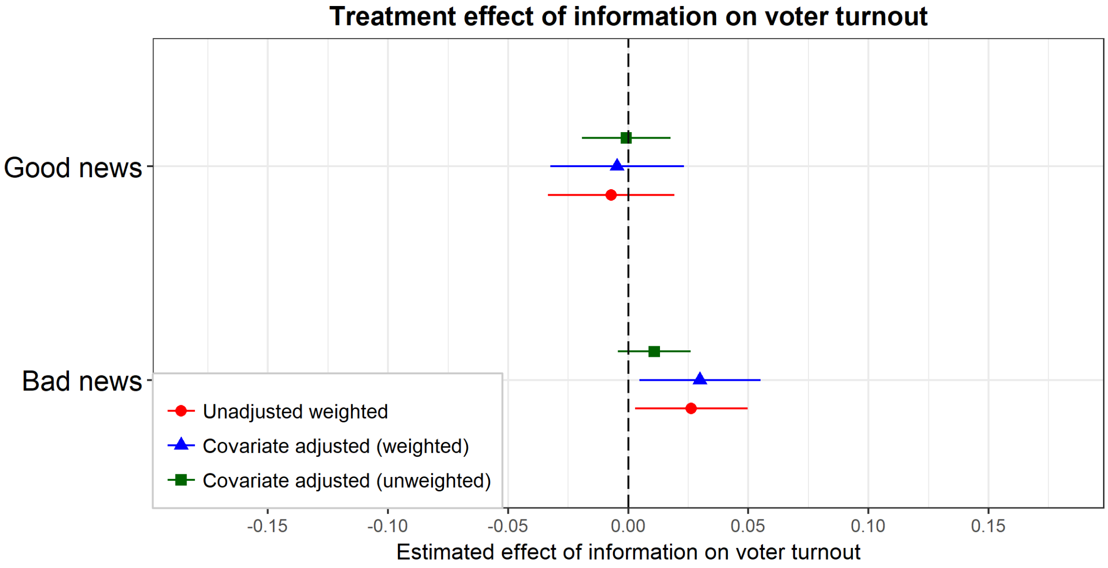

Figure 11.2: Estimated change in the proportion of voters who turn out to vote receiving good news (top row) or bad news (bottow row) about the politician, compared to receiving no information. The top two estimates in each row are the pre-specified covariate adjusted analyses, with a full set of covariate-treatment interactions (country-weighted analysis in green squares, unweighted in blue triangles); the bottom row shows unadjusted, weighted estimates (red circles). Horizontal lines show 95% confidence intervals for the estimated change. In all cases, the differences are close to zero and statistically insignificant.

Finally, to test the hypothesis that information effects are stronger when the gap between voters' prior beliefs about candidates and the information provided is larger, the final two columns of Table 11.1 present the results of estimating equation (11.3) on the pooled data

[21] The confidence intervals are slightly tighter in Figure 11.2 than for vote choice in Figure 11.1. This reflects both differences in the variances of the outcome variables–0.56 in the unweighted average for the 0-1 vote choice variable vs. 0.81 for turnout, and thus greater variance for the former—and differences in the Ns, with more data on self-reported turnout. See Appendix tables for details.

set (including both the good and bad news groups).[22] Overall, the average causal effect of information is indistinguishable from zero—and we find no evidence that the magnitude of the impact depends on the gap between voters' prior beliefs and the provided information.[23]

Table 11.1: Effect of Information on Vote Choice and Turnout

	Good news		Bad news		Pooling good and bad news	
	Vote choice	Turnout	Vote choice	Turnout	Vote choice	Turnout
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.014	−0.005	−0.004	0.030*	0.006	0.012
	(0.017)	(0.014)	(0.018)	(0.013)	(0.013)	(0.010)
N_{ij}	−0.018	−0.006	−0.062***	−0.010	−0.060***	0.003
	(0.017)	(0.014)	(0.016)	(0.013)	(0.012)	(0.011)
Treatment * N_{ij}	0.001	0.011	−0.0001	0.028	0.008	0.004
	(0.022)	(0.019)	(0.021)	(0.015)	(0.013)	(0.011)
Control mean	0.444	0.848	0.454	0.839	0.437	0.844
RI p-value	0.058	0.456	0.851	0.78	0.099	0.991
Covariates	Yes	Yes	Yes	Yes	Yes	Yes
Observations	6,355	8,630	7,687	10,152	14,112	18,850
R^2	0.411	0.206	0.295	0.153	0.327	0.149

Note: Columns 1-4 estimate equations (11.1) and (11.2), while columns 5-6 estimate equation (11.3). "Vote choice" indicates support for the incumbent candidate or party. Standard errors are clustered at the level of treatment assignment. We include block fixed effects and a full set of covariate-treatment interactions. Control mean is the unadjusted average in the control group. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

[22] Note that in the Mexico study, the measure of N_{ij} is defined at the randomization block level; so that study is not included in regressions that include N_{ij}, since the measure of N_{ij} is colinear with the block fixed effects. Similarly, in Brazil, there is only variation in $N_i j$ in the bad news case, given that in the case of municipal account rejections, the information is dichotomous (rejected or not) and good news that confirms positive priors is treated the same as good news that leads to updating.

[23] Table 11.1 shows results with imputation of block-level averages for missing covariate values, as pre-specified; the Online Appendix shows very similar results without imputation.

11.1.3 Consistency of Results Across Studies

How do the findings of particular individual studies compare to our overall results? Despite our focus on harmonization, there are nonetheless important differences across the common interventions in our studies. Indeed, prior to fielding these projects, we considered the ways in which we expected these differences to condition treatment effects.

Consider, for example, that performance information was attributed alternately to candidates or to parties, depending on whether the electoral system makes one or the other type of cue more pertinent. In Mexico, mayoral term limits (with no immediate reelection of incumbent candidates) make information on the performance of individual candidates less relevant; in Burkina Faso, closed-list proportional representation (PR) makes party cues most salient.[24] In Benin, despite a multi-member district system, particular incumbents are associated with and understood to "represent" particular geographic areas (communes) within multi-member constituencies.[25] In other settings—such as Ugandan general elections or Indian state assembly elections—it could be feasible to use either party or candidate cues, as each candidate is associated with one party but also represents single-member constituencies, and those studies opted for information about candidates.[26] There are other distinctions across studies, for example, in the office of politicians (e.g., mayor or member of parliament), the type of performance information provided, and the medium for communicating the information. As we discussed in Chapter 3, these contextual differences are important—and could, in principle, account for any differences in results across the settings in which our interventions were fielded.

Strikingly, however, we in fact find negligible differences across studies in the effects of

[24] See Arias et al. (Chapter 5) and Lierl and Holmlund (Chapter 8) for evidence on the relevance to voters of information about party performance.

[25] See Adida et al. (Chapter 4) for evidence.

[26] See Platas and Raffler's (Chapter 6) study of candidates for Ugandan Parliament, Buntaine et al.'s (Chapter 7) study of Ugandan district councilors and district council chairs, or Sircar and Chauchard's study of state assembly elections in the Indian state of Bihar (Chapter 10).

the common informational treatment. In Figure 11.3, we compare the effects of the common intervention arm across countries and studies. Here we focus on unadjusted analyses; however, substantive results are similar for covariate-adjusted country-specific analysis. Note that for reasons discussed further below, these estimates may differ slightly from those presented in earlier chapters—for example, because of differences in the analysis protocol that was pre-specified for individual chapters and that pre-specified for this meta-analysis. These differences are mostly minor, however, and the figure therefore provides a useful summary of findings in Part II of the book.[27]

As Figure 11.3 shows, not only is the overall meta-analysis result indistinguishable from zero—but the estimates for every single country, and for both good and bad news, are statistically insignificant as well. Given these results, the heterogeneity across contexts and interventions could be viewed as an inferential advantage: despite this variation, the effects of information appear quite similar—and quite uniformly weak —across diverse settings. The figure also underscores the potential value of pooling studies in a meta-analysis, however, since the confidence intervals in several studies are rather wide (owing to the fact, inter alia, that only a portion of subjects were allocated to the common intervention or control conditions, while others were included in alternate treatment arms). Pooling the studies gives us greater precision and greater confidence in the robustness of the null findings.

We find similar results for turnout in Figure 11.4. Not only is the average effect overall a tightly estimated null, as previously suggested in Figure 11.2, but also the estimated effect is null in each study. Overall, our findings for vote choice and turnout are striking: not only is there no evidence for any effect overall, but there is almost no evidence for an effect using the pre-specified meta-analysis in any of our six completed studies.

[27] See Chapters 2 and 3.

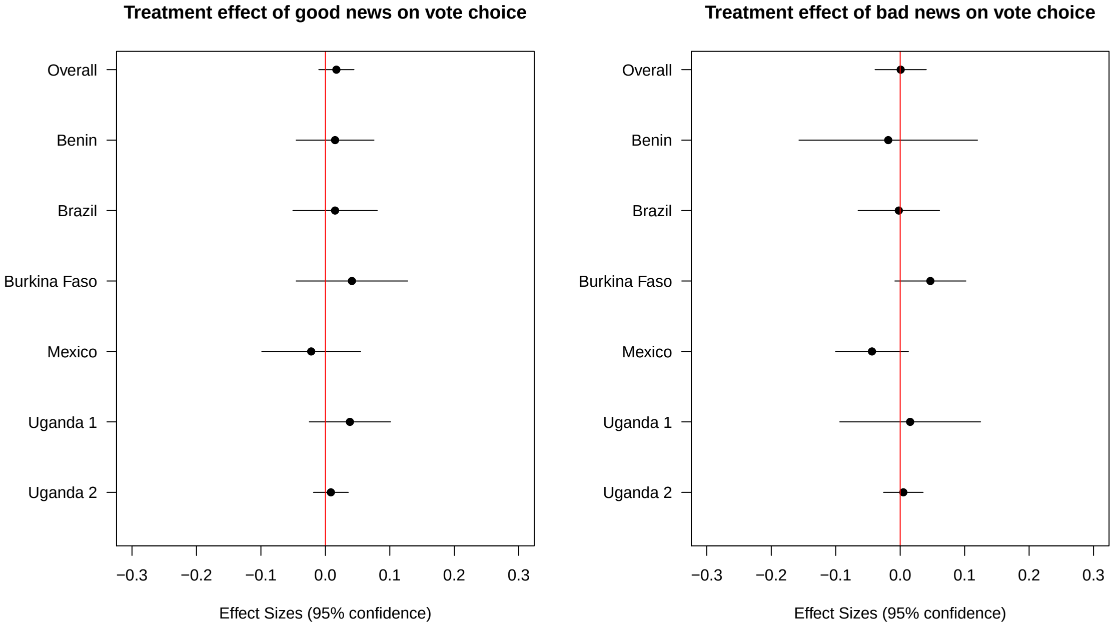

Figure 11.3: Estimated change in the proportion of voters who support an incumbent after receiving good news (red circles) or bad news (blue triangles) about the politician, compared to receiving no information. Weighted unadjusted estimates. Horizontal lines show 95% confidence intervals for the estimated change. In all cases, the differences are close to zero and statistically insignificant.

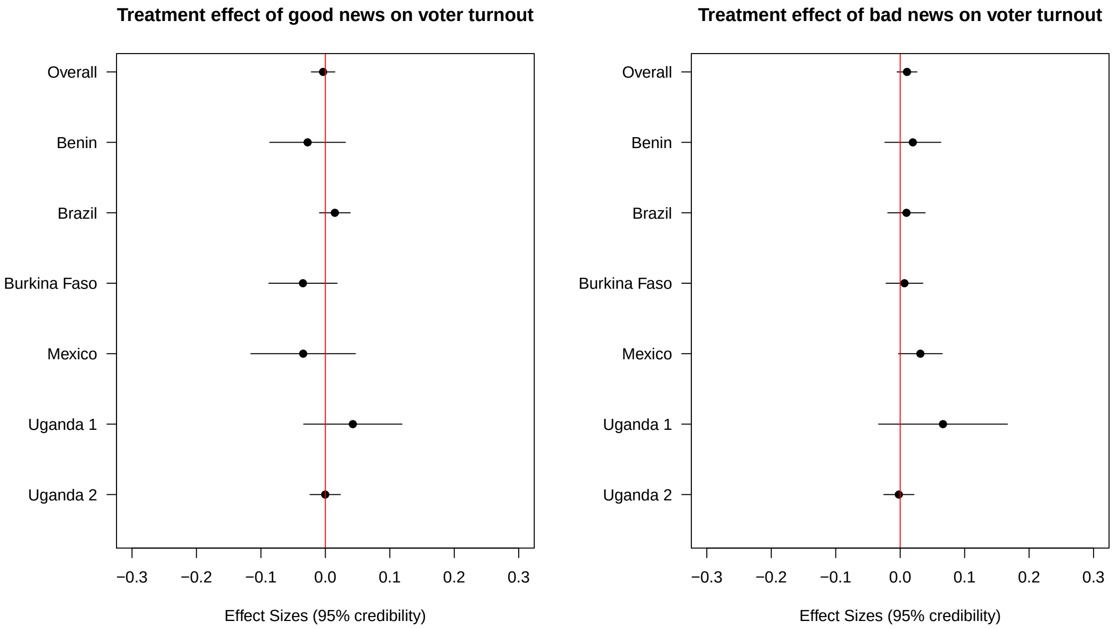

Figure 11.4: Estimated change in the proportion of voters who turn out to vote after receiving good news (red circles) or bad news (blue triangles) about the politician, compared to receiving no information. Weighted unadjusted estimates; substantive results are similar for covariate-adjusted country-specific analysis [see Online Appendix]. Horizontal lines show 95% confidence intervals for the estimated change. In all but one case, the differences are close to zero and statistically insignificant.

11.2 Secondary Analysis: A Bayesian Approach

An alternative approach to meta-analysis takes as the target of inference a general parameter associated with a class of processes, rather than the average effect in a set of cases.

Here we implement such an analysis, similar to that pre-specified in our MPAP as a secondary analysis, following the approach used by Rubin and others in the analysis of the effects of training on student performance in eight schools.[28] For a general treatment of this example and approach see Gelman and colleagues; see also our discussion in Chapter 2, section 2.3.5 (this volume).[29]

The key feature of the approach is that we assume that the treatment effect in a particular case, i, is drawn from a population of treatment effects with mean μ and standard deviation τ. Note that there is no imposition of homogeneity across cases—of homogeneity for that matter. If indeed there is large fundamental heterogeneity then we should infer a large τ. Note also that "fundamental" uncertainty here does not mean that common logics do not obtain across places; it is possible that heterogeneity arises because of other unmodeled features, such as characteristics of subjects or of polities. If modeled, the mean μ could be a function of these features, and we would expect lower values of τ. Given the lack of observed heterogeneity in effects, we do not pursue that approach, however.

The simplest analysis, which we present here, uses only the information provided above on the estimated effects and estimates of uncertainty (clustered standard errors) for each case, which we will call $\widehat{\mu}_j$ and σ_j. We place flat priors on μ and on τ (subject to a non negativity constraint), and the likelihood function uses the probability of observing the estimate for a given

[28] In the MPAP we specified an analysis that assesses the distribution of effects based on the count of votes for the incumbent and the total number of voters. The analysis as specified however is at odds with the design since it does not take account fo the fact that the treatment was randomized within blocks. Accounting for this would require a more complex multilevel structure with block and country effects; instead we elected to use a closely related model that is similar in spirit but that uses the study level estimated effects as inputs.

[29] Gelman et al. 2013. For a nice informal introduction to this approach see the guest post on Andrew Gelman's Blog https://tinyurl.com/eight-schools.

country $\widehat{\mu}_j$ given σ_j and parameter μ_j, where the probability of μ_j is itself a function of μ and τ:

$$\mu_j \sim N(\mu, \tau)$$

$$\widehat{\mu}_j \sim N(\mu_j, \sigma_j)$$

Note that this analysis treats the individual case estimates as if they were drawn from a common distribution. This is clearly a very strong assumption and requires at a minimum a conceptualization of the kinds of cases that form the population as well as an assumption that the selection of a case is not related to the size of the treatment effect in the case. In addition the particular model assumes normality; this is also a substantive assumption though not as fundamental as the assumption regarding case selection. See also our discussion in Chapter 2, section 2.3.

Bayesian analysis allows for estimation of the parameters of this model: μ, τ and $\mu_j, j = 1, 2, \dots 6$. The results are shown in Figure 11.5 for candidate support for the good news and bad news cases, and Figure 11.6 for turnout.

We see from these results that the estimated μ is very similar to the estimated average effect in our main frequentist analyses, in all cases very close to zero. We also estimate quite a low level of fundamental heterogeneity, which in general spans zero. Finally, as is typical in such models, we see that our individual estimates for cases are in general closer to our estimate of μ than the estimates generated by each case separately. Note that exceptional cases—for instance, the larger point estimates of good news for the Uganda 1 and Burkina Faso studies—get substantially revised in this meta-analysis, reflecting the singularity of the results but also the fact that they are themselves measured with considerable uncertainty.

To further probe the robustness of this result, we also conducted an analysis in which we sequentially leave out one study and estimate μ and τ under this assumption (not shown). The

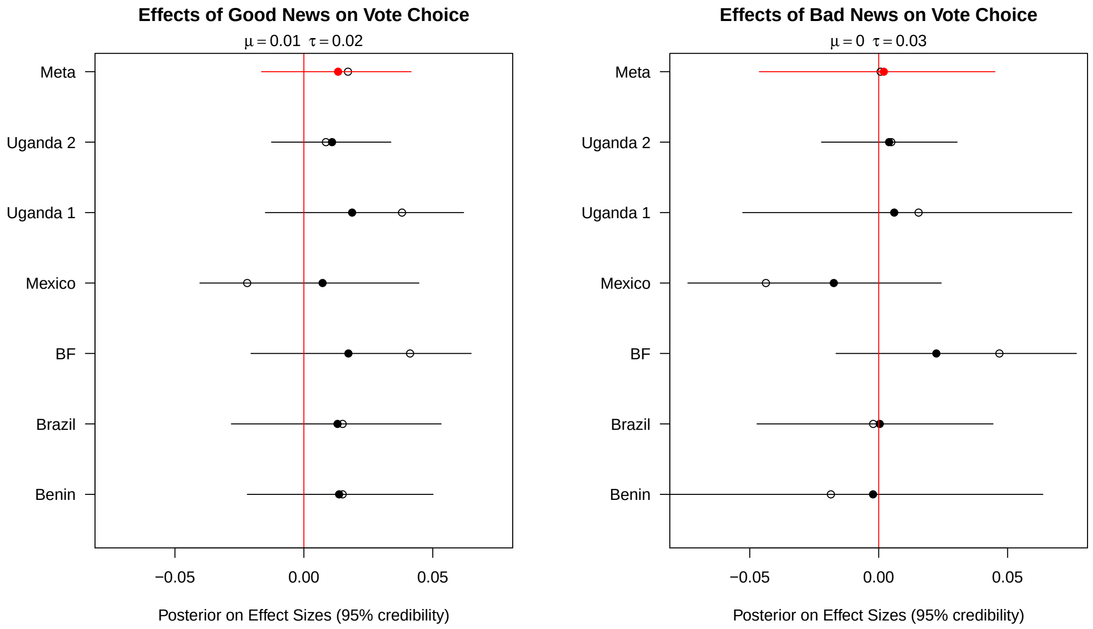

Figure 11.5: Vote choice estimates from Bayesian model. The solid dots and lines show the estimates from the Bayesian model; the top row shows the overall meta-estimate of μ and τ. The white dots show the original frequentist estimates: in many cases shrinkage can be observed, especially in cases that have effects that are more imprecisely estimated.

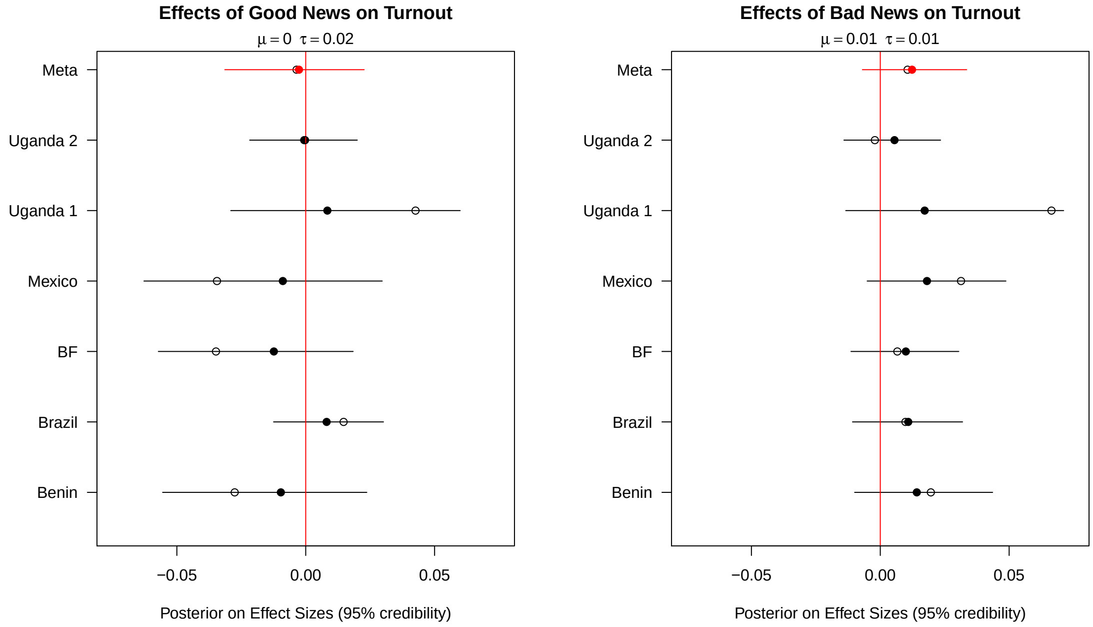

Figure 11.6: Turnout estimates from Bayesian model. The solid dots and lines show the estimates from the Bayesian model; the top row shows the overall meta-estimate of μ and τ. The white dots show the original frequentist estimates: in many cases shrinkage can be observed, especially in cases that have effects that are more imprecisely estimated.

analysis confirms that overall results differ little from those in Figure 11.5 and 11.6.[30]

Overall, the Bayesian results support the conclusion of our frequentist analysis: effects of the common intervention are small, and quite uniformly small, across cases.

11.3 Robustness of Results

How robust are these null results? Two possible threats to the validity of our conclusions bear special scrutiny: (1) reliability of our outcome measures; and (2) study-level attrition. We discuss both of these potential concerns here, focusing on our primary analysis. We conclude that these likely do not undermine our substantive conclusions. We also discuss (3) further differences between the results presented in individual chapters of Part II and the analysis of country-specific results using our meta-analysis specifications, to assess the extent to which pre-specified modeling and data analysis choices could be driving our findings.

11.3.1 Reliability of outcome data

A first consideration involves our outcome data, in particular, the contrast between self-reported vote choice and aggregate official results. Biases in the self-reported data may certainly exist; see for instance the comparison of these types of data in Adida et al. (this volume, Chapter 4) or Arias et al. (this volume, Chapter 5).

However, it is not probable that reporting unreliability of the individual-level data explain our null results. After all, social desirability-type concerns might suggest that voters in the treatment group would differentially over-report vote choice for incumbents, at least in the good news group. This conjecture however might lead us to falsely reject true null hypotheses—rather than fail to reject false nulls. We also draw from a number of studies that used secret-ballot measures of self-reported vote choice, and which found self-reported voting outcomes

[30] See the Online Appendix.

that substantially track official results; see e.g. Boas et el. (Chapter 9) on Brazil or Lierl and Holmlund (Chapter 6) on Burkina Faso. In these studies with quite good self-reported voting data, estimated effects of information are also null. Finally, where studies can rely on official returns, for example, in estimating aggregate effects at the level of polling stations, we find results that are broadly consistent with those we report in this chapter.[31]

Turning to secondary outcomes, our study teams also measured individual-level turnout. To be sure, mean reported turnout is fairly high, at 85% in the pooled control group in Table 11.1, which may reflect social desirability bias. Yet we might expect this to operate symmetrically across the treatment and control groups, or, as with vote choice, to lead to overreporting of turnout among treated respondents, at least in the good news group. The bias would thus run against the null findings. Given the high level of self-reported turnout, ceiling effects could conceivable account for the weak effects. Even so, there appears to be room for movement—and yet Table 11.1 and Figure 11.2 show quite precisely estimated null effects.

11.3.2 The missing India study

Second, could study-level attrition account for our null overall results? One advantage of our pre-specification of studies and of integrated publication is that it makes implementation failures—and missing studies—evident. This is an advantage from the point of view of transparency. Yet of course, missing studies also limit our ability to draw inferences to the whole study group. Our planned India study did not occur due to local political backlash, as Sircar and Chauchard (Chapter 10) describe and as we discuss in section 11.4.2. If politicians correctly anticipated large effects of our informational interventions in that context—and in consequence moved to block implementation of the study—this could indicate that treatment effects would have been larger in India, had the study in fact occurred.

Whether the dropping of the India study leads to bias in estimates of the overall treatment

[31] See e.g. Chapters 4 and 5.

effect depends ultimately on unknowables. Our MPAP noted the following: "If there is attrition for entire blocks containing four or more treatment and control units (for example if entire studies fail to complete or if regions within countries become inaccessible) these blocks will be dropped from analysis without adjustment unless there is substantive reason to believe the attrition is due to treatment status." On the one hand, as Sircar and Chauchard (Chapter 10) detail, the planned India study did not occur due to logistical and implementation problems in one treatment village, which was somewhat atypical in that local politicians came from a small, independent party; Sircar and Chauchard had negotiated agreement to their study with all of the largest parties in Bihar, but not with that party. On the other hand, given the presence of this small party in other villages, it is also plausible that the exposure of any of those villages to treatment would also have resulted in study-level attrition. In other words, India could have dropped out of the study under almost any possible treatment assignment vector. In that case, study level attrition would not be related to treatment status.

Of course, such conjectures are ultimately unverifiable. It is therefore useful to conduct a sensitivity analysis in which we ask the following question: how big (in absolute value) would the estimated effect in India have to have been to have produced a non-null estimated effect in the overall meta-analysis, given findings with our other six studies?

We can answer this question with some algebra. Let $\widehat{\mu}$ be the unweighted average estimated effect in the six realized studies, $\widehat{\theta}$ be the estimated effect in India had the study taken place, and $\widehat{\gamma}$ be the average effect we would have estimated had all seven studies taken place. Then the unweighted average effect across the seven studies is

$$\widehat{\gamma} = (6\widehat{\mu} + \widehat{\theta})/7, \tag{11.4}$$

and its standard error is

$$\sigma_{\widehat{\gamma}} = \sqrt{36\sigma^2_{\widehat{\mu}} + \sigma^2_{\widehat{\theta}}}/7, \tag{11.5}$$

where $\sigma^2_{\widehat{\mu}}$ is the variance of $\widehat{\mu}$ and $\sigma^2_{\widehat{\theta}}$ is the variance of $\widehat{\theta}$.[32] Then the t-stat for the estimated average treatment effect across the 7 studies would have been greater than 1.96 if and only if the estimated effect in India had satisfied the following inequality:[33]

$$\widehat{\theta} \geq 1.96\sqrt{36\sigma^2_{\widehat{\mu}} + \sigma^2_{\widehat{\theta}}} - 6\widehat{\mu}. \tag{11.6}$$

First assume an SE of 0.016 in India (that is, 1.6 percentage points for the 0-1 vote choice variable); this is the smallest of the study-specific standard errors seen in our baseline specifications.[34] Note that this assumption is likely to be conservative, since the India study clustered treatment assignment at the polling station level. Considering only the common intervention arm and the control group, there were to be 400 polling stations with 20 citizen respondents in each polling station; see Chapter 10 and the India team's pre-analysis plan. This implies that in the good news case, we would have needed an estimated average treatment effect of 0.094, or 9.4 percentage points, to see a significant effect in the seven-study meta-analysis.[35] To place an additional bound on this estimate, we can perform the same calculation inputting the largest country-specific standard error (0.071).[36] Under this assumption about the variance associated with the India study, we would have needed an estimated average treatment effect of 0.136—that is, 13.6 percentage points—for the seven-study meta-analysis to register a finding statistically distinguishable from zero.[37] That is a very, very large effect—much bigger than anything we see

[32] This is because $\text{Var}(\widehat{\gamma}) = \text{Var}[\frac{6\widehat{\mu}+\widehat{\theta}}{7}] = \frac{36\text{Var}(\widehat{\mu})+\text{Var}(\widehat{\theta})}{49}$, and the square root is the standard error. This calculation assumes independence of the effect estimates across countries. We took many measures to ensure that results in one study would not affect others—for example, by blinding researchers to results in other studies until all studies had been completed.

[33] The t-stat is given by $\widehat{\gamma}/\sigma_{\widehat{\gamma}} = 6\widehat{\mu} + \widehat{\theta}/\sqrt{36\sigma^2_{\widehat{\mu}} + \sigma^2_{\widehat{\theta}}}$.

[34] See the Online Appendix.

[35] We have the largest point estimate in absolute value in the good news case with the vote choice outcome, so using this example is the hardest test for the claim that seeing results in India would not alter our overall conclusions.

[36] Online appendix.

[37] We are grateful to Fredrik Sävje for his advice on this approach.

in other studies, including those, like Mexico, where we also saw evidence of political responses to the treatment implementation. It therefore appears highly unlikely that completion of the India study would have altered our overall conclusions.

11.3.3 Country-specific analyses vs. Meta-analysis

There are several salient ways in which the country-specific results presented in this chapter differ from those presented in the chapters of Part II, even with respect to analyses of the common intervention arm. We do not see this as problematic, especially since most of those differences were pre-specified. Different studies can approach the same data in different ways, and sometimes even reach different substantive conclusions. Yet, clear pre-specification of the different approaches allows readers to see in a transparent way what distinctions may be driving the different findings.

Nonetheless, it is worth considering whether these differences could affect our findings—and whether we would reach different conclusions if we aggregated the data in some other way. Some of the most significant differences arise with the analysis of data from Mexico. As Arias et al. discuss in Chapter 5, a baseline survey was prohibitively expensive in that study; thus, rather than use individual-level prior perceptions of incumbent malfeasance as the measure of P, the authors estimate the randomization block-level average from questions in the endline survey, using only control-group respondents.[38] After gathering individual-level outcome data (e.g. vote-choice and turnout) in the control group, they show the treatment flyers to control-group respondents, and ask again about perceptions of malfeasance of the incumbent party. Finally, they use the change in perceptions from prior to posterior to operationalize good and bad news (see Chapter 5).

From the perspective of the meta-analysis, however, this approach has several disadvantages.

[38] See Arias et al., Chapter 5, for discussion of the assumptions necessary for this approach to recover the average priors in the treatment group; essentially, no within-block spillovers and inter temporal stability of perceptions are the key elements, together with randomization of the treatment.

Most importantly, it is based on the updating of perceptions rather than the performance information (Q) itself. In addition, it is necessarily defined at the randomization block level; the analysis in Chapter 5 largely focuses on precinct-level analysis of official electoral returns, though the common individual-level, self-reported vote choice and turnout measures are also analyzed. In our main analyses reported in the previous subsection, we therefore operationalize good and bad news in Mexico using an alternate approach. First, we examine the distribution of the difference between the two percentages presented on the flyer shown to the treatment group in the common intervention arm: that is, the percentage of unaccounted or misspent funds in the subject's municipality, and the percentage in the other municipalities in the state governed by opposition parties. Where that difference is positive, it indicates that corruption was greater in the respondent's municipality, whereas negative differences indicate greater possible corruption in opposition municipalities. Following our MPAP's definition of good and bad news, which defines good news in part in relation to the median of the distribution of the performance score in the relevant comparison group, we then define respondents to have received good news if they receive a below-median difference between the percentages and bad news if they receive an above-median difference.[39]

However, it is important to consider how results would change if we use other definitions of good and bad news. Using the definition in Chapter 5, but individual-level outcome data, we find no substantive difference in our meta-analysis results but, oddly, a strongly negative effect of *good* news for Mexico; this result is also reported and discussed in Chapter 5 and its Online Appendix. Using an alternate definition that subtracts the *individual*-level prior from an

[39] An alternative, not as tightly linked to our MPAP, would take positive differences as bad news and negative differences as bad news; however, this leads to a much larger bad news than good news group, because incumbent municipalities governed by the PRI tended to perform worse. Note also that we do not attempt to relate this measure of Q to a block-level measure of P for two reasons: (1) P is measured on a different scale, namely, a five-point scale of perception of incumbent malfeasant spending, whereas Q is continuous measure of percentage differences; (2) block-level priors (as measured by perceptions in the control group at the start of the endline survey) are empirically almost completely non-predictive of Q, so subtracting block-level P from block-level Q gives no improvement in statistical efficiency. The measure of Q we use is a straightforward alternative that is closely linked to the spirit of our MPAP.

individual-level posterior, measured in both the treatment and the controls groups, we do not find this negative effect, and the overall meta-analysis results are similar in our main specification. However, in one pooled specification using that measure that includes LCV councilors as well as chairs from Uganda 2 study (see discussion later in this subsection), we see a marginally significant positive pooled effect of good news. Yet, using this individual-level measure of the difference between the posterior and the prior to define the good and bad news groups risks post-treatment bias; indeed, we find treatment assignment predicts the prior belief in both the good and bad news groups, leading us to discard that measure. In sum, in one specification there is a hint of the possibility of modest effects of good—but not bad—news. Overall, though, the weak effects are remarkably stable to the different ways of operationalizing good and bad news in the Arias et al. study.

Several other differences between the country-specific analyses in this chapter and those in Part II are worth further mention as well. The study by Boas et al. (Chapter 9) uses a pre-specified Lasso routine to select covariates, while here we use those specified in the MPAP (that were gathered consistently across studies). As we showed, however, we find substantively similar results with unadjusted and covariate-adjusted analyses. Our analysis of Uganda 1 data (Platas and Raffler, Chapter 6) uses data only on incumbents, as pre-specified in the MPAP; Platas and Raffler (Chapter 6) find somewhat more evidence of effects when looking at the performance of opposition candidates in their “meet the candidate” debates, especially when restricting analysis to credible candidates who ended up winning a minimum percentage of the vote. Finally, and perhaps most importantly, the Uganda 2 study of Buntaine et al. (Chapter 7) finds heterogenous effects across type of office, with significant effects of information on vote choice when looking at LCV councilors, but not LCV chairs or LCIII councilors or chairs. In the main meta-analysis, we focus on LCV chairs yet we conduct additional analyses pooling both LCV chairs and councilors (and clustering standard errors by respondent, since in these analysis two vote choice outcome variables are recorded for each respondent). However, in our main analyses, we find no significant treatment effects in the pooled meta-analysis, whether we include or exclude LCV councilors.[40]

Overall, then, our results are remarkably consistent to different ways of operationalizing the good and bad news groups, different measures of the outcome variable, and different sub-groups of the population. We turn further to heterogeneous effects in the next section, yet here we find that our weak results are strikingly consistent across specifications.

11.4 Making Sense of The Null Findings

What explains the weak effect of information on voter behavior in our pooled data?

Figure 3.1 in Chapter 3 outlined a causal chain through which informational interventions might shape vote choice, and ultimately political accountability.[41] According to this framework, existing information must be disseminated, and it must be received and understood by voters. Those voters in turn must update their perceptions or beliefs in response to the new information. This updating must then produce changes in their voting behavior, ultimately leading them to sanction poorly performing politicians or reward well performing ones. As discussed in Chapter 3, this is the route through which adverse selection—the choosing of "bad types"—can be reduced and thus political accountability can be improved.

However, there are numerous ways such a causal chain can break down. Any one of the nodes may not be operative, undermining the links and connections that lead ultimately to electoral sanctioning. Voters may not understand or pay attention to the information. Even if they do, the information may not lead to any updating of perceptions, particularly if the information is not salient or from a credible source. Or, voters may believe the information and their perceptions may change as a result of it; yet their voting behavior may be so fully

[40] See, however, the previous comment in relation to our discussion the Mexico study, that we find a marginally significant effect of good news in one specification of the meta-analysis that uses a measure of good and bad news defined with individual-level posteriors, and that includes both councilors and chairs from Uganda 1.

[41] See also Lieberman, Posner and Tsai (2014) or Kumar, Post and Ray (2017)

determined by partisanship, ethnic relations to politicians, or clientelism—among other factors—that the information makes no marginal difference to their voting. Other possibilities, including feedback loops not countenanced by the simple diagram in Figure 3.1, might also explain null effects of information. If politicians are aware of the information dissemination, they might seek to counteract it—for example, by increasing vote buying or disrupting the dissemination of information in treated areas, thereby weakening treatment effects.[42]

In this section, we use our pooled data to assess these possibilities, particularly by testing the hypotheses about intermediate outcomes registered in our MPAP (see the book's appendix). We also use observational and experimentally induced variation to evaluate both what may be driving the overall null effect and what alternative forms of information dissemination might have had stronger effects than those we found. We note that while we endeavored to measure all of the variables registered in the MPAP in a symmetric and consistent fashion across studies, this was not always possible, or it did not always take place. In some cases, as considered in the MPAP, this was because not all the variables are interpretable in the same way in each context; for example, co-ethnicity has very different implications and meaning in, say, Benin or Uganda than it does in Mexico or Brazil. Co-ethnicity data were not gathered in the Mexico case for this reason (especially in the states in which the studies were fielded, the concept is not salient). In other cases, teams simply did not gather all the data, or they ended up for various reasons asking questions in non-standard ways that unfortunately did not allow ready pooling across cases. In our analyses below, we therefore pool results for a particular intermediate outcome or conditioning variable using only the countries for which data on the relevant indicator were gathered. In many of the tables, we present country-by-country results, but only for those studies where comparable data were gathered.

[42] See Cruz, Keefer and Labonne (2017).

11.4.1 Voter Perceptions: Reception of Information and Updating

In each of our cases, third-party information on politician or party performance existed; and it was successfully disseminated by researchers and the third-party organizations with whom they partnered, in the sense that the flyers, SMS messages, videos, and other experimental stimuli were in fact fielded and directed to voters in the treatment groups.

One possibility, however, is that treated respondents simply did not on average absorb the information to which they were exposed. Table 11.2 assesses this possibility, using manipulation checks that are specific to each study. The table shows that overall, respondents did pass these manipulation checks. The findings are driven by the Mexico and Uganda 1 studies. In Uganda 1, as reported in Chapter 6, exposure to good news (though not bad news) had a large, statistically significant and positive effect on respondents' expectations: those exposed to videos conveying good news about a candidate increased their expectation that the candidate would exert more effort than other candidates if elected. Platas and Raffler also found that watching their debates increased respondents' political knowledge, measured as a index of the roles and responsibilities of Members of Parliament, share of candidates known, and knowledge of candidates' policy priorities. Arias et al. (Chapter 5) find that good news leads Mexican voters to favorably update their beliefs about the integrity of the incumbent party, while bad news leads voters to update unfavorably (though the estimate of the latter effect is not statistically significant). In Table 11.2, we assess the effects of treatment assignment on an indicator variable for whether the respondent correctly recalled the content of the flyer. One caveat, however, answers to this question may not be clearly interpretable for respondents in the control group (who did not receive any flyer). We also therefore explored whether assignment to the treatment made respondents significantly more likely to remember receiving such a flyer; we find that it did, with 6 percent of the control group and 32 percent of the treatment group stating that they remember the flyer, a difference of 26 percentage points that is highly statistically significant (t-statistic is 16; not shown). We note also that in the analysis protocol we are using here, the manipulation check is not significant in Brazil; yet using their block by treatment interactions in Chapter 9, Boas et al. show significant effects of treatment on knowledge of whether accounts were accepted or rejected. Overall, simple failure to absorb the information—or to "receive" it, in the language of the causal chain in Chapter 3, Table 3.1—does not fully explain the null results, since we saw null effects of the informational treatments even in those cases with strong evidence of respondent comprehension of the information.

Nonetheless, the non-results in three of the five cases in the table suggest the substantial difficulties involved in disseminating information information that voters receive and understand. Some of the contrasts between the cases could be due to the dissemination technology; for example, the SMS messages deployed by Uganda 2 can be a difficult way to convey nuanced messages.[43] It is nonetheless remarkable not to see more evidence on the manipulation check from contexts in which respondents were shown graphical presentation of information.[44] As we discuss elsewhere, this difficulty appears a critical practical challenge for organizations that would like to increase political accountability through informational interventions.

Moreover, even if there is evidence that overall and across studies, information was communicated and a portion of voters received it, this does not imply that their perceptions changed as a result of it.[45] We registered two hypotheses about politician characteristics that we believed might change through the provision of performance information (the numbering here as elsewhere follows our MPAP):

- H3: Positive (negative) information increases (decreases) voter beliefs in candidate integrity.
- H4: Positive (negative) information increases (decreases) voter beliefs that candidate is

[43] Fafchamps and Minten (2012); Aker, Collier and Vicente (2017)

[44] Consider, for example, the case of the Brazil study, which distributes audit information very similar to Ferraz and Finan (2008), albeit via direct delivery at the individual level rather than the natural dissemination at the municipal level via community radio featured in that study.

[45] See step 4 in the chain in Figure 3.1.

Table 11.2: Effect of treatment on correct recollection, pooling good and bad news [unregistered analysis]

	Manipulation check: Correct Recollection					
	Overall	Benin	Brazil	Mexico	Uganda 1	Uganda 2
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.060^{***}	-0.038	0.038	0.166^{***}	0.146^{***}	0.001
	(0.015)	(0.038)	(0.021)	(0.023)	(0.034)	(0.010)
Covariates	No	No	No	No	No	No
Observations	15,755	897	1,677	1,898	750	10,533
R^2	0.316	0.320	0.378	0.151	0.075	0.245

Notes: The table reports results on manipulation checks across studies, using recollection or accuracy tests at endline that were specific to the content of each study's interventions (MPAP measure M30). The dependent variable, correct recollection, is dichotomized in each study using the following measures: Benin: whether correctly recalled the relative performance of incumbent in plenary and committee work; Brazil: whether correctly recalled whether municipal account was accepted or rejected; Mexico: identification of content of the flyer; Uganda 1: index consisting of knowledge of MP responsibilities, MP priorities for constituency, and identities of contesting candidates. Individuals with an index equal to or greater than 1.5 on a 0-3 scale were coded as correct recalls; Uganda 2: whether correctly recalled relative financial accountability relative to other districts.*p<0.05; **p<0.01; ***p<0.001.

hardworking.

Table 11.3 reports results for our pooled measures of politicians' integrity and effort, respectively. We measure perceptions of incumbents' integrity and effort using similar questions across studies.[46] Estimates of the effect of information is statistically indistinguishable from zero in both the good and bad news groups, as well as in the whole study group. We also show in the

[46] Sample question on MPAP measure M5 of candidate effort: "In your opinion, does [INCUMBENT] make much more, a little more, a little less or much less effort to get things done than other deputies in this [Department]?" Sample question from MPAP measure M6 of candidate integrity/honesty: "How surprised would you be to hear from a credible source about corruption involving your [MP/Mayor/Councilor]? Would you say you would be (1) Very surprised (2) Somewhat surprised (3) Not too surprised (4) Not surprised at all."

Online Appendix that information does not in the aggregate change the *importance* that respondents attach to different policy priorities, such as community and personal benefits, politician efficiency and integrity, or ethnic or partisan identity. Note that there is considerable scope of learning, as we showed in Chapter 3, in that correlations between our aggregate measures of priors (P) and politician quality (Q) are present but also modest; prior beliefs, however, are linked to perceptions of other candidate characteristics and to vote choice. However, here we find no overall impact of the information on perceptions of politicians' characteristics, at least on these dimensions. We consider later, in our discussion of heterogeneous effects, possible reasons for the finding that voters overall absorbed the information and yet posteriors over candidates on the dimension of the information may not have budged. For instance, we consider there the question of whether voters filter the information through partisan lenses.

Table 11.3: Effect of information on perception of importance of politician effort and honesty

	Good news		Bad news	
	Effort	Honesty	Effort	Honesty
	(1)	(2)	(3)	(4)
Treatment effect	-0.008	-0.061	0.084	0.102
	(0.047)	(0.052)	(0.054)	(0.105)
Control mean	2.545	2.754	2.299	2.732
Covariates	No	No	No	No
Observations	7,056	7,053	7,743	8,765
R^2	0.270	0.274	0.354	0.220

Note: The table reports results pooling Benin, Burkina Faso, Brazil, Mexico, and Burkina Faso. MPAP measures M5 (effort) and M6 (honesty). Regressions include randomization block fixed effects; standard errors are clustered at the level of treatment assignment. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Overall, then, an important conclusion is that one major breakdown of the causal chain in Figure 3.1 is at steps 3 and 4: voters received and assimilated the information, but only substantially so in two cases; and in most cases, the disseminated information did not cause them to update their perceptions of candidate effort and honesty. Why not? One possibility is that the information was not provided by a credible source. Of course, perceptions of the source credibility could vary both across studies and for different individuals in the same study. We measured perceptions of the credibility of different possible sources of information.[47] We can therefore code whether the information source deemed most credible by a particular respondent was in fact the source of the information to which she was exposed (or would have been exposed, if in control) in the study in which she was included.

Table 11.4 presents an exploratory analysis, which we emphasize was not pre-registered; our goal in presenting it is to assess whether source credibility and the treatment interact, looking at perceptions of effort and integrity as outcomes. We find no evidence here that the credibility of the information source interacts with treatment. That is, at least as measured here, the treatment is not significantly more effective when the respondent has deemed its source to be credible, among the options given in MPAP measure M24.[48]

Given the lack of apparent connection between the informational treatments and perceptions of politician effort and honesty, it is also useful to assess how those perceptions in turn correlate with vote choice. We emphasize that such an analysis does not shed any light on the causal effect of those perceptions on electoral support; nor does it tell us whether any influence of information on perceptions would in turn lead to an impact on vote choice. Nonetheless, it is interesting to see that in the unregistered analysis in Table 11.5, there is a strong significant association between perception of the incumbents' effort and honesty and voters' subsequent electoral support for the incumbent.

[47] The sample question for M24 in the MPAP reads as follows: "Suppose that you received information about a politician, for example, information about how he or she had performed in office. Which of the following sources would you trust the most [second most; third most] for that information? [READ OPTIONS]: (a) Local politician; (b) Flyer or pamphlet from an NGO; (c) A person conducting a survey; (d) An influential member of your community; (e) In a debate between candidates; (f) Other."

[48] See previous note.

Table 11.4: Effect of information and source credibility on evaluation of politician effort and honesty [unregistered analysis]

	Incumbent vote choice			
	Good news		Bad news	
	Effort	Honesty	Effort	Honesty
	(1)	(2)	(3)	(4)
Treatment	0.050	−0.042	0.116	−0.042
	(0.074)	(0.089)	(0.083)	(0.089)
Credible Source	0.069	−0.085	0.012	−0.085
	(0.075)	(0.069)	(0.074)	(0.069)
Treatment * Credible Source	−0.140	0.020	−0.041	0.020
	(0.091)	(0.095)	(0.098)	(0.095)
Control mean	2.545	2.744	2.291	2.744
Covariates	No	No	No	No
Observations	6,406	6,239	7,052	6,239
R^2	0.280	0.300	0.350	0.300

Note: Regressions include randomization block fixed effects; standard errors are clustered at the level of treatment assignment. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Overall, the evidence thus far supports the idea that the breakdown in the information and accountability chain occurred both at the level of reception and especially the perception of the information. Observational evidence may suggest that had perceptions been altered, vote choice might have been influenced as well. It is difficult thus far to say why the interventions had little impact on respondents' updating; but we return to that question later.

11.4.2 Politician Response

We registered another hypothesis that may bear on the connections between information and accountability along the causal chain. Perhaps politicians respond to negative information by

Table 11.5: Relationship between evaluation of politician effort and honesty with vote choice

	Incumbent vote choice			
	Good news		Bad news	
	(1)	(2)	(3)	(4)
Effort	-0.061^{***}		-0.055^{***}	
	(0.007)		(0.009)	
Honesty		-0.032^{***}		-0.042^{***}
		(0.006)		(0.006)
Covariates	No	No	No	No
Observations	5,011	4,896	5,881	6,412
R^2	0.510	0.447	0.353	0.332

Note: Regressions include randomization block fixed effects; standard errors are clustered at the level of treatment assignment. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

altering their campaign strategies. Politicians have a menu of options to counterbalance "bad" information: they can divert more time to campaigning in treatment areas, they can increase vote buying, and they can counteract negative impacts of the information by undermining the credibility of the information source.[49] At the extreme, they may attempt to stop the dissemination efforts altogether.

We pre-registered this hypothesis as

- H5: Politicians mount campaigns to respond to negative information.

Indeed, we see substantial evidence that politicians were not passive and in some cases indeed attempted to derail information dissemination efforts. Sircar and Chaucard (Chapter 10), for example, describe how the actions of representatives of a small party in Bihar, India imperiled the safety of some of their enumerators and ultimately led to the termination of their fieldwork.

[49] See Cruz, Keefer and Labonne (2017), Humphreys and Weinstein (2013).

Arias et al. (Chapter 5) describe similar episodes in several municipalities in Mexico. There, incidents included not only potential threats to enumerator safety but also the fabrication by political actors of fake fliers; the fliers were designed to mimic those distributed by the research team's NGO partner, but unlike the real fliers provided explicitly partisan negative information. These episodes did not, however, lead to the cancellation of the project in the Mexican case. On the other hand, see also Platas and Raffler (this volume) on politicians' positive reaction to their interventions in Uganda, and Buntaine et al., (this volume) on how the method of dissemination (e.g., SMS) can affect politicians' ability to counteract negative information.

We can assess quantitative evidence for backlash to some extent as well. Research teams in the Benin, Brazil and Mexico projects asked treatment and control group respondents a question similar to the following: "In the week before the election did you hear of [incumbent candidate] or someone from their party making statements about [the dimension of information provided to treated groups]?"[50] As pre-specified, we account for the clustered nature of treatment assignment when comparing treated and control respondents—and the presumably clustered nature of politicians' response, in targeting treated areas. As Table 11.6 shows, treatment had a substantial and statistically significant effect, elevating "yes" responses to the question about incumbent statements by 7 percentage points overall, with significant effects individually in the Mexico study (but not Benin). Following H5, we focus only the bad news case.

Yet, can such politician response explain our null effects? Probably not, for several reasons. First, it appears unlikely that this backlash occurred as systematically as would be required to counteract a true effect of the information interventions on voters. In Mexico, for example, we find quantitatively that treatment did provoke politicians' backlash, and have qualitative evidence on attempts to prevent our intervention in a handful of municipalities.[51] However,

[50] Measure M8 in our MPAP. The question was not included in the Brazil, Burkina Faso, or Uganda 2 instruments; and the India study did not complete an endline survey. We have data on this question for Uganda 1, but treatment is assigned at the individual level, complicating the assessment of politician backlash—which is presumably targeted at particular areas and which would therefore affect both treatment and control individuals in those areas.

[51] This is parallel to the situation in India, where one village caused the problems that led to the stopping

Table 11.6: Effect of bad news on politician backlash

	Politician response / backlash		
	Overall	Benin	Mexico
	(1)	(2)	(3)
Treatment effect	0.070*	0.068	0.070**
	(0.030)	(0.057)	(0.023)
Control mean	0.108	0.068	0.146
Covariates	No	No	No
Observations	2,052	702	1,350
R^2	0.578	0.504	0.848

Note: Dependent variable measures perception of incumbent campaigning on dimension of disseminated information. Backlash measured for studies with clustered assignment. Standard errors clustered at the level of treatment assignment. Regressions include randomization block fixed effects; standard errors are clustered at the level of treatment assignment. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

while politician's backlash was positively correlated with the amount of malfeasance reported in the fliers, it was not correlated with whether voters interpreted the information as good or bad news. In other words, the response of politicians did not take into account voters' prior beliefs. Also, such a hypothesis would also not be consistent with the null effect of good news we find even in those settings where backlash did not occur.[52] Finally, we estimate null effects even in those contexts, like Benin, where we have no qualitative or quantitative evidence of politician backlash.

A more plausible hypothesis may therefore be that interventions providing positive or negative performance information in fact have little impact on voters—yet politicians often believe

of implementation.

[52] We presume here that politicians did not systematically mount campaigns where information was on average favorable; note that they could have shifted greater attention toward control areas, which could also diminish our ability to identify a positive treatment effect for good news.

that they will. In many contexts, politicians misjudge the preferences and behaviors of their constituents, and they may therefore misjudge the impact of information about their performance on voters.[53] Politicians may also tend to react because they are risk averse, especially given the high cost of campaigning, and especially where levels of political competition are relatively high. As we noted in Chapter 1, our interventions focus on the selection mechanism: they are targeted at voters, whose sanctioning is key in many models of political accountability. They were not designed, however, to address the moral hazard (politician) dimension. Relatedly, the timing of the interventions may be important: is there sufficient time for the information to become part of the larger campaign debate? One worry in delivering information so shortly before election is that it gives enough time for the incumbent to punch back, but not enough time for challengers to counter-punch/hammer home the information being provided.[54] Differences in timing of the intervention relative to the election could conceivably underlie the different findings of the well-known Ferraz and Finan study—which found very large impacts of publicizing corruption allegations in Brazil, but in the year before an election—and the findings reported by Boas, Hidalgo, and Melo (Chapter 9).[55] As we discuss in the conclusion to this chapter and the conclusion of the book, such hypotheses generated by our findings are interesting and should be explored in greater depth in future research.

11.4.3 Learning From Variation

In this section, we test the remainder of the hypotheses from our MPAP. In addition to accumulating evidence on the average effects of our interventions across studies, our Metaketa also aimed to learn from various kinds of context- and intervention-specific variation across the studies. We

[53] See for instance Broockman and Skovron (Forthcoming) on the extent to which politicians misjudge voter ideology, or Rosenzweig (2017) on the extent to which they overestimate the efficacy of electoral violence.

[54] See, for example, Grossman and Michelitch (forthcoming) on the importance of the timing of information campaigns to the options available for politicians' responses.

[55] Ferraz and Finan (2008).

considered three types of variation in our MPAP: (1) heterogeneous treatment effects across characteristics of the respondents; (2) variation in contexts and features of interventions; and (3) experimentally induced variation, for instance, through the inclusion of alternative treatment arms. Understanding such variation in effects can shed light on the voter types for whom effects are strongest, and allows us to assess whether those types are relatively rare in our population, possibly explaining our overall null effect. It can further point us towards future interventions that might be more effective. And testing these hypotheses allows us to assess further possible breakdowns along the causal chain from information to accountability.

Substitution effects

We pre-specified several conditional hypotheses related to steps 3, 4, and 5 in Figure 3.1 in Chapter 3. First, we expected that information would operate on vote choice in part by reducing the weight voters place on ethnicity, copartisanship, and clientelistic relations. In particular, we expected that good news would reduce the bias for voting against non-coethnic outgroup candidates and bad news to reduce the bias for voting for coethnic candidates.[56] However, even though information may reduce the weight voters place on these relations, we expect that information has more positive effects for voters that do not share ethnic, partisan, or clientelist ties with candidates. We conceptualized these as hypotheses as *substitution effects*, in the sense that ethnicity or partisanship could provide heuristic substitutes for information: in particular, that information effects are more positive for voters that do not share ethnic identities; that information effects are more positive for voters with weaker partisan identities; and that information effects are more positive for voters who have not received clientelistic benefits from any candidate.

These substitution effects were pre-specified as follows:

- H6: Information effects are more positive for voters that do not share ethnic identities.

[56] As we noted in the MPAP, this hypothesis is not relevant for all projects, e.g. Brazil or Mexico where the measurement of coethnicity is either trickier than in other contexts, or irrelevant

- H7: Information effects are more positive for voters with weaker partisan identities.
- H8: Information effects are more positive for voters who have not received clientelistic benefits from any candidate.

We can assess such hypotheses by looking at sub-group effects within and across studies. We emphasize, however, that a causal interpretation of these heterogeneous effects is not justified by the experimental design. We do not manipulate the conditioning covariates in our experiments, and we lack an identification strategy that would allow us to make strong causal claims about the effects of these variables. Nonetheless, comparing and contrasting effects in different sub-groups could in principle give important hints about mechanisms that can explain our findings.

Table 11.7 presents tests of hypotheses H6–8. Here, we interact treatment exposure with our measures of co-ethnicity, co-partisanship, and clientelism. We emphasize that here we can back out heterogeneous treatment effects, i.e., we can assess whether the effect of treatment differs among the incumbent's co-ethnics or co-partisans, but we do not experimentally identify the effects of co-ethnicity or co-partisanship on vote choice. Nonetheless, the associations between these variables and vote choice are noteworthy. Interestingly, co-ethnicity is not strongly associated with vote choice in these data, though this may be due in part to the inclusion of a case, Brazil, in which co-ethnicity is not highly salient for vote choice.[57] Additionally, in those contexts where co-ethnicity is expected to be salient, the lack of association may be due to lack of within-block variation in co-ethnic relations between voters and politicians: voters living in the same villages tend to share ethnic relations. (We use fixed effects for randomization blocks in these regressions, as in all of our regressions). Co-partisanship, however, is very strongly associated with vote choice, in these regressions resulting in a nearly 20 percentage point increase in the probability of voting for the candidate. Our measure of clientelism, meanwhile, has a negative and significant association with vote choice; though the sign appears odd at first glance, it may reflect the way in which the question was asked, as discussed momentarily. These

[57] See Bueno and Dunning (2017).

observational associations help to validate the measures of both the moderators and vote choice, a topic we assess further later as well.

Most importantly for testing our hypotheses on substitution effects, however, we find no evidence that the strength of the treatment varies as predicted by H6-H8. Neither for co-partisanship nor clientelism do we find any evidence of a significant interaction. We note one ambiguity of measurement for co-partisanship, which is that our common indicator actually measures strength of attachment to the incumbent, rather than the strength of partisan identities overall.[58] It is possible that a voter who is not very attached to the incumbent's party has strong attachments to another party—or no partisan attachment at all. However, H7 would predict different effects on average for those who are attached to the incument's party and those who are not, since the latter group plausibly includes both opposition partisans and non-partisans or swing voters. Moreover, in the Online Appendix, we present additional exploratory specifications to test H7, for example, a quadratic specification and a table in which we present treatment effects at each level of scales measuring partisan attachment to the incumbent (rather than dichotomizing co-partisanship as we do in this chapter). As in Table 11.7, we see essentially no evidence that the treatment effect varies with the partisan attachment to the incumbent.

To investigate the potential moderating effect of clientelism, we use responses to the following question, implemented—with minor variations—across all of our study sites: "How likely is it that the incumbent, or someone from their party, will offer something, like food, or a gift, or money, in return for votes in the upcoming election." Responses are recorded on a 4-point scale ranging from "not at all likely" to "very likely." We acknowledge an important flaw with this measure, however. It does not ask respondents to say whether they personally have benefited (or expect to benefit) from a handout from the incumbent—something that would have picked up economic dependency, and thus possible unwillingness to desert the incumbent in response

[58] Here is the sample question for M19, from the MPAP: "On this scale of one to seven, where seven means you are very attached to [INCUMBENT'S PARTY], and one means you are not very attached to [INCUMBENT'S PARTY], what degree of attachment do you feel for [INCUMBENT'S PARTY]?"

to our informational treatments. Rather, it captures respondents' beliefs about how likely the incumbent is to engage in clientelistic mobilization and corrupt vote-buying practices more generally. This should be borne in mind when interpreting the results.

We employ a dichotomous, subjective measure to assess whether the treatment is more effective in cases where respondents and politicians do not share ethnic ties. Specifically, enumerators posed the question: "Thinking of the [incumbent politician], would you say that [you come from the same community/share the same ethnic group/share the same race] as this candidate?" Note that we do not have this measure for Mexico or Burkina Faso since we did not judge ethnicity to be a salient dimension of political identity in these settings. We do see an interaction effect for co-ethnicity, but only in the good news case; and we note in any case that H6 suggests information effects are more positive for non-coethnics. Overall, these data lend little support for these hypotheses about substitution effects.

Context-specific heterogeneity

We considered possible sources of variation in effects across interventions according to the context in which they were delivered. Notably, we expected information to have greater impact in contexts where information was less readily available at baseline, and among voters with less exposure to information in the pre-treatment period.

- H9: Informational effects are stronger in informationally weak environments.
- H10: Informational effects are stronger in more competitive elections.
- H11: Informational effects are stronger in settings in which elections are believed to be free and fair.

To operationalize these tests at the individual level (as opposed to the system level), we asked respondents to state how certain they were about about their priors regarding politicians' performance or background. Our assumption is that voters are uncertain about their priors when they have worse access to information, making this a reasonable proxy.

Political competition is another dimension of variation that might be thought to shape the impact of the treatment. We hypothesize that voters will be more attentive to information—and more willing to devote time and cognitive resources to processing it—in environments where electoral competition is tight, and thus their vote is more likely to be pivotal in swaying the final result. Voters may quickly tune out information in places where the election outcome is a forgone conclusion. Counterarguments also suggest themselves, however. For instance, electorally competitive environments might already be flooded with information—as parties, journalists, and civil society groups typically focus more attention on those races—attenuating the effects of any additional news, of the kind delivered by our interventions. Clientelistic distribution might also be more prevalent in these settings, and thus economic motivations to stick with the incumbent may crowd out the impact of performance-related information. The question of whether competition moderates information's effect is therefore an empirical one.

We measure competition using administrative data. In countries using simple plurality voting, competitiveness is calculated at the constituency level and is given as one minus the margin of victory for the winning candidate—over the runner up—in the most recent election. For countries using proportional representation, the calculation is more involved, and is performed at the party or candidate level, depending on whether the system employed is closed list or open list, respectively.[59]

The third contextual hypothesis we specify relates to election fraud. If voters suspect that their vote will not count—perhaps because they expect politicians to stuff ballot boxes or doctor vote totals—or if they believe their vote choices will be observable to an incumbent who may punish them for voting the "wrong" way, then information interventions may fall flat. To gain empirical traction on these issues, survey teams posed two questions to respondents. First, enumerators asked how likely it is that "powerful people can find out how you vote, even though there is supposed to be a secret ballot in this country." Second, voters were asked whether the

[59] The full description of these variables is provided in the meta-PAP, measure M25; see the Appendix.

counting of votes in the forthcoming election is likely to be free and fair. We interact these ordinal measures—available for individuals—with the treatment indicator and look for evidence of a significant interactive effect.

The results of the tests on H9 and H11 are displayed in Table 11.8. All but one of the six coefficients on the interaction terms are very small in magnitude and are statistically insignificant at conventional levels. There appears to be no firm support for the claim that local contextual factors moderate the effect of the informational interventions. That said, we do estimate a significant interaction between the treatment and our measure of certainty in priors for the bad news sub-group. This negatively signed interaction suggests that the treatment is less effective the more uncertain voters are in their priors about politicians—a highly counterintuitive finding. Digging deeper, we find that this result is driven largely by Benin, where, due to the low reporting of priors, only 161 observations are available for this estimation. The unexpected direction of this result, its appearance in only one of the six studies, and the large number of statistical tests we are running make us hesitate to place much credence in this finding.

In fact, however, we see little evidence of such variation in effects by country. As we have seen, the results are highly consistent—and strikingly weak—in every study in the Metaketa (Figures 11.3 and 11.4). Overall, then, such hypotheses about heterogeneity across the common intervention arms receive little support from our comparative data. To be sure, this is not to say that the perceived welfare relevance of information never conditions the effects of its provision, however. In Adida et al.'s study, for example, information on legislative performance of deputies—absent other information on the civic importance of legislative productivity—may actually have a negative effect, when voters value clientelism or local public goods provision over legislative activity. These particular cross-national differences were not pre-specified. See Chapter 4 and also our discussion of alternative intervention arms below. But overall, it is the consistency of the weak results across studies that is most notable.

Finally, we test H10 with another set of regressions. Because our measures of electoral competitiveness vary at the block level—and our regressions include block fixed effects—we split the samples at the median level of electoral competition and run our block fixed effects regressions. Here, too, we find no evidence for this kind of context-specific heterogeneity driving our results.

Intervention-specific heterogeneity

It is also possible that features of the interventions themselves—and voters attitudes toward them—provide an explanation for the overall null result. We pre-specified three hypotheses:

- H12: Information effects—both positive and negative—are stronger when the gap between voters' prior beliefs about candidates and the information provided is larger.
- H13: Informational effects are stronger the more the information relates directly to individual welfare.
- H14: Informational effects are stronger the more reliable and credible is the information source.

To elaborate, we might expect that quite small gaps between voters' prior beliefs about politicians and the "truth" about those politicians—as revealed by our interventions—may be relatively immaterial for updating and voting behavior. On the other hand, large discrepancies of this sort—indicating that the voter was more seriously misguided in her prior beliefs—may be generate measurable effects on outcomes (H12). We test this hypotheses using individual-level data on the distance between voters' priors and the information provided: that is, the variable N_{ij} (see Chapter 3). We see whether treatment effects become more positive as N_{ij} grows larger.

Next, it might be the case that intervention matters for voters only when it coveys information perceived to be relevant or salient. For instance, in deciding how to cast their vote, some citizens may care little about how often incumbents attend legislative committee meetings, believing instead that a politician's diligence in attending to constituency work is the more important yardstick of performance. Similarly, some citizens may worry deeply about the corruption in public administration, whereas other may view this as a secondary concern. At baseline, the Metaketa teams presented respondents with a list of activities in which their local incumbent politician(s) might regularly be involved. Respondents had to describe which of these activities they would most like to receive information about. We generate a dichotomous variable indicating whether or not the activities that were the subject of the actual intervention—activities that differ across studies—matched the activity described by the respondent as being the one they were most interested in.

Source credibility is a third element of the interventions that might be thought to affect the effectiveness of information. When citizens perceive information to come from a credible source—for instance, a respected NGO, or the national newspaper of record—voters may be more likely to update their beliefs in accordance with the information supplied. By contrast, we anticipate that citizens will tend to discount information originating from less dependable sources (e.g. opposition parties or slanted news sources). Naturally, credibility is likely to be in the eyes of the beholder. We therefore ask each respondent, at baseline, to state which information sources they trusted the most. We then define a binary variable that takes one when the the source ranked as the most trustworthy by respondents is the same as the true information source for the intervention—which, once more, varies across countries.[60]

The results of heterogeneous effects analyses employing these three measures are reported in Table 11.10. In making inferences, we look to see whether the interaction between the treatment indicator and these moderating variables enters the regression as statistically significant. In fact, none of them do. The data reveal no signs that gaps in prior beliefs, information salience, or source credibility moderate the effects of the treatment. Thus we cannot conclude that intervention-specific heterogeneity explains our overall null findings.

[60] In addition, the India study of Sircar and Chauchard (Chapter 10) planned an evaluation of H14 through experimental manipulation of the identity of the messenger, but that intervention was not fully fielded and outcome data were not collected.

Demographic characteristics

Finally, in addition to pre-specified hypotheses about moderators (substitution effects), several of the studies presented in Part II of the book present additional findings on heterogeneous effects. For instance, in Benin, Adida et al. (Chapter 4) find evidence of stronger effects among younger and poorer voters.[61] In Mexico, Arias et al. (Chapter 5) find that treatments mattered more in high-competition as well as in low-information environments (the latter measured as places where voters were more knowledgeable about politics or had higher levels of media consumption); in the Uganda 1 study, Platas and Raffler (Chapter 7) find that the good news treatment mattered among those who thought debates were a credible source of information and among those who expected favors from the politician if he were elected; and in the Uganda 2 study, Buntaine et al. (Chapter 8) find that good news significantly increased vote share for LCV councilors and bad news significantly decreased the vote share for LCV councilors; the interventions had no effects on LCV chairs, or LCIII chairs and councilors.[62] Some of these findings have been assessed previously in this chapter using pooled data; and not all of these hypotheses are testable in the pooled meta-analysis, given the smaller number of covariates for which data were collected. However, several of them are. Thus, having been derived inductively in those cases, hypotheses about these sub-group effects can then be tested on the whole dataset.

In Tables 11.11 and 11.12, we present the results of estimating full interaction models, showing the estimated coefficients on interactions between treatment and covariates as well as the constituent terms. Note that some of the covariates were assessed previously in this chapter but here we present the full regression as specified our MPAP. The first two columns show the estimates for the pooled metadata (the second column includes both LCV chairs and councilors in the Uganda 2 study, while the first column excludes councilors). The other

[61] They also find evidence, for their alternative arms, that effects were strongest among those who received the worst news; see Chapter 4 for discussion.

[62] LCV is a district-level local council in Uganda, while LCIII is a sub-county local council the jurisdiction of which is nested within LCVs.

columns then show country-specific regressions (there are again two columns for Uganda 2, with the first of those excluding LCV councilors). We only include covariates that were measured in comparable ways across all studies. As with our previous analysis of the gap between priors and information, here we see some associations between the covariates and votes for the incumbent candidate/party about whom information was provided. For example, wealth and previous support for the incumbent are positively and significantly associated with incumbent vote choice; so, interestingly, is exposure to clientelism, but also belief that the vote is secret and elections are free and fair. These associations are not the focus of our conditional hypotheses, however; rather, we seek to assess the heterogeneity of treatment effects across values of these covariates.

As indicated by the general lack of significance of the interaction terms, here we find little evidence, at least per the linear interaction model, that treatment effects vary conditional on these covariates. We do see some evidence in particular countries. The Metaketa approach provides a very useful way to test sub-group effects derived from one country on a wider dataset. In this case, however, we see no confirmation of the existence of such heterogeneous effects.

11.5 Does Public Information Boost Informational Effects? The Impact of Alternative Interventions

Overall, we find null effects of our common intervention arm; and very little apparent variation in effects across different forms of contextual, intervention-specific, and individual-level heterogeneity. It seems most plausible that breakdown in the information and accountability chain occurred due to the fact that voters did not, in the main, receive and especially update their perceptions as a function of the information.

However, the structure of the Metaketa also was intended to allow assessment of alternative interventions that might prove more effective than the common intervention arm. Finally, we sought to explore divergent effects within studies—especially from experimentally induced variation in the delivery of treatments. In particular, we forecast that comparisons between the common and alternate intervention arms within each study might provide insights into the conditions under which information was more or less effective.

The studies in Part II of this book report intriguing evidence in this regard. For example, Adida et al. (Chapter 4) find that the common intervention focused on legislative performance did not have significant average effects on electoral behavior; but they show evidence that treatment works when it is combined with (1) a civics message educating people about the welfare importance of legislative productivity; and (2) the information is widely disseminated in lots of villages in a constituency. Platas and Raffler find that publicly screened videos (rather than the privately screened videos associated with the common intervention) increased political knowledge and slightly but discernibly affected vote choice in Uganda (Chapter 7). And Boas, Hidalgo, and Melo used the second arm of their field experiment to inform voters about municipal-level changes in scores on the National Literacy Evaluation during the mayor's first term. Among parents of children enrolled in school, for whom the issue should be most salient, they find that voters punish poor performance and reward (or are indifferent to) good performance. They conclude that a personal connection to the policy in question may be a prerequisite for information about incumbent performance to change voting behavior.

Such hypotheses are interesting and promising, and should be tested systematically. While we cannot evaluate all of them in this Metaketa, we fortuitously had three projects with similar alternative arms, in which information was provided to voters in a public rather than private fashion. As underscored by the project-specific PAPs for those projects, the hypothesis was that the provision of information in a public rather than private setting would generate common knowledge of the intervention and foment greater collective action—and therefore evidence a greater impact on vote choice. We also registered this hypothesis in the MPAP.[63]

- H15: Informational effects are stronger when information is provided in public settings.

We pool data from the three projects with public treatment arms to assess this hypothesis. Table 11.13 reports the pooled effect as well as the effect in each country. Here we regress vote choice for the incumbent on an indicator for the private information condition and an indicator for the public information condition. As anticipated by the analysis in this chapter, we estimate a null effect on the private condition—but a large and statistically significant effect of the informational treatment in the public condition. This is driven by an extremely large effect in Benin.

11.6 Chapter Conclusion

The meta-analysis presented in this chapter suggests that informational interventions, at least of the kind we have considered in this research project, are not an effective way of driving voter behavior. In a well-powered meta-analysis pooling data from six of seven planned experimental studies on the impact of performance information, we find no evidence of apparent impact on vote choice or electoral participation. Stronger than an insignificant meta-result, we find that no individual country displays significant impacts of voter information interventions conducted shortly before the election. Neither the directionality of the information shock (good versus bad news) nor the magnitude of the shock (difference from priors) generates changes in voter behavior. Turnout, as well as choices among voters, is unaffected. Measures of the ways that voters feel about their politicians are similarly unchanged overall, suggesting that while perceptions may be important drivers of voting behavior, none of the types of intervention studied

[63] We cannot fully assess one remaining hypothesis in the MPAP systematically: H16: Informational effects are not driven by Hawthorne effects. We discussed the possiblity of randomizing the content of consent forms but did not implement this across the studies; in part, our commitment to informed consent in all cases limited our capacity to estimate its effect through comparison to randomized control groups that did not receive the consent request.

here were meaningful drivers even of perceptions. None of the forms of heterogeneity to which we pre-committed are present in the data, and sub-group effects reported by individual studies do not manifest themselves as general meta-results. Hence, while transparency-promoting interventions appear to represent a potentially attractive way of deepening democracy, we find no evidence that they are effective.

What do we learn from this meta-result that is different from what could be gleaned from any individual study? First, of course, there is the issue of power: any individual study if powered normally has a 20% chance of failing to find a result that is actually there, while our meta-study will have a much lower probability of Type II error. By replicating a non-result in six contexts, we can conclude with a degree of statistical certainty that would not otherwise be possible. There is also an important point about implementation to be made. When looking at any individual study, there is always the question as to whether implementation on the ground was simply so weak as to have failed to test the hypothesis at all. The aggregation of six studies—none of which had major obvious implementation problems of this kind—makes it much less likely that our lack of results are arising from something that simply went wrong on the ground. The lack of meta-impacts even on perceptions of politician performance suggest a set of important foundational questions for future research: how performance in specific dimensions is incorporated into an overall perception of politician quality, and the way that the credibility of the information source may alter the degree of updating.

Stepping back, these results speak to the comparative impact of transparency-promotion interventions more broadly. As discussed in Chapter 3, our studies all sought to manipulate only the selection margin of voter choice; taking place immediately before the elections they were not intended to induce an incentive effect on politician behavior. However, because the most obvious mechanism generating pressure on politicians to respond is precisely the effectiveness of information on the selection margin, our non-result should imply that politicians have no reason to respond to such interventions at all. In this case these programs would similarly not have generated an improvement in politician moral hazard even if they had been introduced further before the election. Of course, the fact that in two of our studies politicians attempted to end or undermine the intervention suggests that in some cases they did in fact perceive it as a threat, and introduces the possibility that we have lost from the study precisely those circumstances under which the information would have been most important. Normatively, politicians should have the opportunity to respond to information and defend themselves against particular charges of malfeasance or ineptitude.[64] Yet, the fact that they may do so is of more than academic concern, given that real-world implementers would face similarly heterogeneous opposition to implementation by political leaders. The implication is that informational interventions can only be easily conducted in contexts where they will be ineffective. In this sense, our findings provide important information to donor collaboratives, policy makers and project implementers. In light of the interest in and excitement among such organizations about using informational interventions with voters to boost transparency and accountability, our core results are perhaps depressing, but important.

At the same time, our results do point to interesting alternative conditions under which information may have more impact—in particular, our pooled findings on the public intervention arms. These and other results that are idiosyncratic to studies reported in Part II, should be assessed systematically, perhaps in future Metaketas. To justify the case for extending this model to other areas, however, it is imperative to have more evidence on the usefulness of the approach itself. It is to this topic that we turn in the next chapter.

[64] For example, criminal charges—while officially recognized—may be politically motivated (India), or politicians' lack of effort in some areas—say, shirking their legislative responsibilities in Benin—may be more than compensated for by efforts in other areas.

Table 11.7: Effect of moderators on incumbent vote choice

	Incumbent vote choice					
	Good news			Bad news		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.020	0.004	0.008	0.0002	0.013	0.001
	(0.016)	(0.032)	(0.015)	(0.026)	(0.028)	(0.020)
Coethnicity	−0.033			−0.005		
	(0.032)			(0.040)		
Treatment * Coethnicity	0.082*			−0.030		
	(0.038)			(0.048)		
Copartisanship		0.196***			0.294***	
		(0.041)			(0.038)	
Treatment * Copartisanship		−0.002			0.013	
		(0.048)			(0.045)	
Clientelism			−0.030**			−0.052***
			(0.010)			(0.012)
Treatment * Clientelism			0.003			0.024
			(0.014)			(0.017)
Control mean	0.461	0.459	0.452	0.485	0.454	0.434
Covariates	No	No	No	No	No	No
Observations	5,546	5,742	6,379	5,705	6,326	7,467
R^2	0.419	0.398	0.391	0.255	0.294	0.265

Note: The following cases are included in each regression: Co-ethnicity—Benin, Brazil, Uganda 1, Uganda 2; Co-partisanship—Benin, Brazil, Mexico, Uganda 1, Uganda 2; Clientelism—Benin, Burkina Faso, Brazil, Mexico, Uganda 1, Uganda 2. Regressions include randomization block fixed effects; standard errors are clustered at the level of treatment assignment. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 11.8: Effect of information and context heterogenity on incumbent vote choice

	Incumbent vote choice					
	Good news			Bad news		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	−0.033	0.020	0.012	−0.032	0.00002	−0.003
	(0.057)	(0.026)	(0.038)	(0.061)	(0.034)	(0.035)
Certainty	−0.020			0.009		
	(0.016)			(0.021)		
Treatment * Certainty	0.018			0.008		
	(0.025)			(0.026)		
Secret ballot		−0.004			0.006	
		(0.008)			(0.012)	
Treatment * Secret ballot		−0.002			0.002	
		(0.011)			(0.014)	
Free, fair election			0.003			0.002
			(0.009)			(0.010)
Treatment * Free, fair election			0.003			0.004
			(0.011)			(0.012)
Control means	0.441	0.449	0.444	0.46	0.438	0.443
Covariates	No	No	No	No	No	No
Observations	5,156	6,562	6,343	5,991	7,798	7,637
R^2	0.447	0.378	0.385	0.313	0.239	0.245

Note: *p<0.05; **p<0.01; ***p<0.001

Table 11.9: Effect of information and electoral competition on vote choice

	Incumbent vote choice			
	Low competition		High competition	
	Good news	Bad news	Good news	Bad news
	(1)	(2)	(3)	(4)
Treatment	-0.001	-0.039	0.017	-0.005
	(0.020)	(0.035)	(0.029)	(0.044)
Control mean	0.401	0.46	0.373	0.393
Covariates	No	No	No	No
Observations	1,308	1,265	1,026	1,113
R^2	0.259	0.214	0.298	0.115

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Table 11.10: Effect of information and intervention-specific heterogenity on vote choice

	Incumbent vote choice					
	Good news			Bad news		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.013	0.028	−0.002	−0.009	−0.030	−0.003
	(0.017)	(0.025)	(0.021)	(0.019)	(0.039)	(0.027)
N_{ij}	−0.021			-0.066^{***}		
	(0.016)			(0.016)		
Treatment * N_{ij}	−0.004			−0.003		
	(0.021)			(0.020)		
Information salient		−0.005			−0.055	
		(0.030)			(0.037)	
Treatment * Information salient		−0.016			0.077	
		(0.037)			(0.046)	
Credible source			−0.003			0.001
			(0.031)			(0.030)
Treatment * Credible source			0.021			−0.018
			(0.035)			(0.038)
Control mean	0.446	0.453	0.453	0.453	0.485	0.44
Covariates	No	No	No	No	No	No
Observations	6,409	5,608	6,051	7,712	5,991	7,152
R^2	0.391	0.419	0.373	0.257	0.251	0.245

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Table 11.11: Interaction analysis: Effect of good news on incumbent vote choice

	Incumbent vote choice, good news						
	ALL	BEN	BRZ	BF	MEX	UG 1	UG 2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment	0.016	0.001	0.006	0.019	0.0003	0.034	0.005
	(0.017)	(0.065)	(0.030)	(0.031)	(0.039)	(0.042)	(0.013)
N_{ij}	−0.028	0.024		0.012		−0.119***	0.006
	(0.015)	(0.060)	(0.000)	(0.024)	(0.000)	(0.023)	(0.009)
Treatment * N_{ij}	0.004	0.193**	−0.020	0.007	0.027	−0.007	−0.008
	(0.010)	(0.069)	(0.022)	(0.032)	(0.019)	(0.019)	(0.006)
Age	−0.001	−0.008	0.001	0.003	−0.006**	0.004*	0.001
	(0.001)	(0.005)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
Treatment * Age	−0.004	−0.071	−0.052*	−0.058**	0.043*	−0.025	0.012
	(0.010)	(0.076)	(0.023)	(0.022)	(0.021)	(0.023)	(0.007)
Education	−0.003	−0.011	0.010	0.008	−0.012	−0.008	0.001
	(0.003)	(0.011)	(0.006)	(0.010)	(0.006)	(0.008)	(0.002)
Treatment * Education	0.007	−0.037		0.010		0.040	0.007
	(0.019)	(0.068)	(0.000)	(0.031)	(0.000)	(0.037)	(0.012)
Wealth	0.025*	0.021	0.068	−0.006	0.017	0.064*	0.011
	(0.012)	(0.048)	(0.036)	(0.023)	(0.028)	(0.032)	(0.008)
Treatment * Wealth	0.002	0.013	−0.004	−0.002	0.005*	−0.004	−0.0003
	(0.001)	(0.008)	(0.003)	(0.003)	(0.003)	(0.003)	(0.001)
Voted previously	0.015	−0.062	0.056	0.073	0.061	−0.072	0.043
	(0.033)	(0.065)	(0.072)	(0.051)	(0.087)	(0.080)	(0.032)
Treatment * Voted previously	0.005	0.017	−0.009	−0.034*	0.004	0.017	−0.004
	(0.005)	(0.017)	(0.008)	(0.017)	(0.011)	(0.012)	(0.003)
Supported incumbent	0.176***	−0.034	0.284***	0.278***	0.319***	0.035	−0.051*
	(0.034)	(0.124)	(0.055)	(0.084)	(0.049)	(0.072)	(0.025)
Treatment * Supported incumbent	−0.040*	−0.097	0.023	0.040	−0.094*	−0.144**	−0.012
	(0.018)	(0.089)	(0.051)	(0.034)	(0.042)	(0.053)	(0.011)
Clientelism	−0.030*	−0.046	−0.068***	−0.068	−0.028	−0.008	−0.002
	(0.012)	(0.073)	(0.020)	(0.056)	(0.038)	(0.023)	(0.006)
Treatment * Clientelism	−0.075	0.053	0.111	−0.049	−0.245*	−0.185	−0.041
	(0.045)	(0.141)	(0.108)	(0.073)	(0.110)	(0.121)	(0.043)
Credible source	−0.001	−0.125	0.025	−0.087	−0.013	0.001	0.057
	(0.037)	(0.160)	(0.105)	(0.056)	(0.089)	(0.061)	(0.037)
Treatment * Credible source	0.033	−0.036	0.094	−0.014	0.221**	0.085	0.082*
	(0.048)	(0.139)	(0.073)	(0.110)	(0.080)	(0.097)	(0.035)
Secret ballot	−0.006	0.100	−0.024	0.106	−0.023	−0.018	0.003
	(0.016)	(0.121)	(0.027)	(0.081)	(0.048)	(0.030)	(0.008)
Treatment * Secret ballot	0.045	0.373	−0.037	0.041	0.095	0.035	−0.037
	(0.048)	(0.237)	(0.137)	(0.085)	(0.104)	(0.083)	(0.050)
Free, fair election	−0.001	−0.176*	0.008	0.022	−0.041	0.033	0.007
	(0.013)	(0.068)	(0.030)	(0.043)	(0.031)	(0.027)	(0.008)
Treatment * free, fair election	0.020	−0.053	0.026	0.105***	0.002	0.053	−0.0002
	(0.013)	(0.088)	(0.026)	(0.031)	(0.037)	(0.032)	(0.009)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	6,804	191	859	735	540	377	4,102
R^2	0.375	0.435	0.462	0.355	0.291	0.159	0.446
Note:					*p<0.05; **p<0.01; ***p<0.001		

Note: The table presents results from fitting equation **??**. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 11.12: Interaction analysis: Effect of bad news on incumbent vote choice

	Incumbent vote choice, bad news						
	ALL	BEN	BRZ	BF	MEX	UG 1	UG 2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment	−0.009	−0.154	−0.022	0.064	-0.039^{*}	0.024	0.006
	(0.017)	(0.084)	(0.030)	(0.033)	(0.020)	(0.040)	(0.016)
N_{ij}	-0.057^{***}	−0.106	-0.100^{***}	−0.001		-0.101^{***}	0.001
	(0.017)	(0.069)	(0.028)	(0.025)	(0.000)	(0.030)	(0.011)
Treatment * N_{ij}	−0.009	−0.185	−0.004	0.022	0.051^{***}	−0.029	−0.001
	(0.010)	(0.097)	(0.021)	(0.036)	(0.015)	(0.021)	(0.007)
Age	0.0001	−0.005	0.00003	0.003	−0.001	0.003	0.001
	(0.001)	(0.004)	(0.002)	(0.002)	(0.001)	(0.002)	(0.001)
Treatment * Age	0.002	−0.066	0.001	−0.014	0.005	0.007	0.012
	(0.009)	(0.058)	(0.026)	(0.023)	(0.013)	(0.023)	(0.007)
Education	−0.003	−0.002	−0.001	−0.008	-0.013^{**}	0.001	-0.006^{*}
	(0.003)	(0.012)	(0.005)	(0.009)	(0.005)	(0.009)	(0.003)
Treatment * Education	−0.013	0.024	−0.063	0.002		0.019	0.028
	(0.021)	(0.078)	(0.034)	(0.035)	(0.000)	(0.040)	(0.014)
Wealth	0.023	−0.016	0.012	−0.024	0.042	0.049	0.026^{**}
	(0.015)	(0.091)	(0.041)	(0.026)	(0.027)	(0.035)	(0.010)
Treatment * Wealth	−0.0004	0.006	−0.001	−0.001	−0.001	−0.005	0.001
	(0.001)	(0.008)	(0.002)	(0.003)	(0.002)	(0.003)	(0.001)
Voted previously	0.012	0.100	−0.053	0.073	0.043	-0.227^{***}	0.001
	(0.039)	(0.271)	(0.089)	(0.053)	(0.046)	(0.062)	(0.032)
Treatment * Voted previously	−0.002	−0.016	0.006	0.003	−0.010	−0.001	0.0002
	(0.005)	(0.020)	(0.007)	(0.014)	(0.006)	(0.011)	(0.004)
Supported incumbent	0.230^{***}	−0.084	0.282^{***}	0.111	0.558^{***}	0.293^{***}	0.077^{*}
	(0.054)	(0.189)	(0.049)	(0.091)	(0.051)	(0.055)	(0.033)
Treatment * Supported incumbent	−0.020	0.060	0.027	0.006	0.026	−0.082	−0.027
	(0.020)	(0.113)	(0.054)	(0.035)	(0.036)	(0.048)	(0.014)
Clientelism	-0.042^{***}	−0.083	-0.086^{***}	0.083	−0.006	−0.021	0.002
	(0.010)	(0.139)	(0.019)	(0.061)	(0.019)	(0.021)	(0.007)
Treatment * Clientelism	0.004	−0.231	0.186	−0.066	0.085	0.145	0.018
	(0.046)	(0.286)	(0.105)	(0.074)	(0.063)	(0.096)	(0.043)
Credible source	−0.012	0.003	0.015	−0.064	0.011	0.007	0.031
	(0.035)	(0.224)	(0.075)	(0.057)	(0.049)	(0.058)	(0.041)
Treatment * Credible source	−0.032	−0.149	−0.052	0.285^{*}	−0.136	-0.171^{*}	0.018
	(0.064)	(0.242)	(0.068)	(0.126)	(0.092)	(0.083)	(0.046)
Secret ballot	0.018	0.010	0.026	0.054	-0.054^{*}	0.008	0.003
	(0.013)	(0.141)	(0.025)	(0.088)	(0.025)	(0.029)	(0.010)
Treatment * Secret ballot	0.022	0.127	0.029	0.183	0.005	−0.047	−0.064
	(0.045)	(0.402)	(0.114)	(0.094)	(0.066)	(0.081)	(0.056)
Free, fair election	0.026^{*}	0.204	0.043	−0.005	-0.046^{*}	0.041	0.003
	(0.013)	(0.235)	(0.027)	(0.053)	(0.020)	(0.026)	(0.009)
Treatment * free, fair election	0.003	0.140	−0.018	−0.020	−0.006	−0.003	−0.008
	(0.013)	(0.153)	(0.029)	(0.033)	(0.022)	(0.033)	(0.010)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	7,308	158	818	707	915	373	4,337
R^2	0.304	0.336	0.420	0.240	0.363	0.306	0.302
Note:							*p<0.05; **p<0.01; ***p<0.001

Note: The table presents results from fitting equation **??**. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 11.13: Private vs Public Information: Effect of good news on incumbent vote choice

	Incumbent vote choice, good news			
	Overall	Benin	Mexico	Uganda 1
	(1)	(2)	(3)	(4)
Private information	−0.011	0.026	−0.033	0.011
	(0.021)	(0.048)	(0.038)	(0.026)
Public information	0.052*	0.135**	0.005	0.025
	(0.021)	(0.045)	(0.038)	(0.022)
Control mean	0.362	0.439	0.489	0.16
F-test p-value	0.008	0.019	0.355	0.618
Covariates	No	No	No	No
Observations	3,719	1,033	1,259	1,427
R^2	0.191	0.171	0.065	0.056
Note:	*p<0.05; **p<0.01; ***p<0.001			

Table 11.14: Private vs Public Information: Effect of good news on incumbent vote choice

	Incumbent vote choice, bad news			
	Overall	Benin	Mexico	Uganda 1
	(1)	(2)	(3)	(4)
Private information	−0.015	−0.027	−0.043	−0.039
	(0.029)	(0.074)	(0.029)	(0.044)
Public information	0.015	−0.042	0.015	0.010
	(0.027)	(0.064)	(0.026)	(0.033)
Control mean	0.445	0.535	0.394	0.467
F-test p-value	0.008	0.019	0.355	0.618
Covariates	No	No	No	No
Observations	3,931	818	2,139	974
R^2	0.116	0.225	0.086	0.121
Note:	*p<0.05; **p<0.01; ***p<0.001			