12

Learning about Cumulative Learning: An Experiment with Policy Practitioners

Gareth Nellis, Thad Dunning, Guy Grossman, Macartan Humphreys, Susan D. Hyde, Craig McIntosh, and Catlan Reardon

12.1 INTRODUCTION

At the heart of the Metaketa initiative is the idea that multiple, coordinated studies allow for stronger inferences about the effect of a treatment or program than standard approaches to research. The previous chapters have shown the model in action. We assessed whether increasing the amount of information available to citizens just before an election causes them to be more likely to turn out to vote, to sanction underperforming incumbents, and to reward high achievers. The studies, both individually and in aggregation, uncover little evidence that informational interventions of this kind influence voter behavior.

In this chapter, we put the Metaketa concept itself to the test.[1] Taking a sample of the intended consumers of the Metaketa's findings as subjects, we investigate (a) whether coordinated research can alter policymakers' beliefs about an intervention's effectiveness, as well as improve their ability to make out-of-sample predictions; and (b) whether results from multiple coordinated research projects have stronger effects than the results of individual, "standalone" studies in this regard.

To date, these questions have mostly escaped systematic investigation. This is somewhat ironic. Champions of experimental methods routinely criticize traditional, nonexperimental impact evaluations for making overly confident claims. However, they have levied this charge without demonstrating that experimental evidence is capable of changing beliefs or decisions on the ground.

[1] Special thanks to Richard Sedlmayr, who was central to conceptualizing this study and encouraged us to do it.

It seems natural to expect that providing policymakers with the best available evidence will lead them to reach better decisions. Yet there are reasons to be skeptical. For one thing, studies rarely speak in perfect unison. Findings can differ across studies owing to contextual factors and sampling variability, both of which complicate the task of drawing general policy lessons. Even when evidence is unambiguous, decision-makers often appear reluctant to change ongoing policies or programs. Reviewing the case literature on how policy is formulated, Weiss and colleagues conclude that "most studies seem to be used in selective bits, reinterpreted to fit existing preferences or ignored."[2] There is a need for a controlled yet realistic test of these issues. The six Metaketa I studies offer a good opportunity to undertake such a test. The topic they address is policy-relevant: information and transparency interventions are a mainstay of democracy promotion efforts worldwide. The study designs are closely congruent with one another, facilitating side-by-side comparison of the findings. Perhaps most importantly, the studies show little heterogeneity: all suggest that the treatment in question has little or no effect. Admittedly, we did not know these results when we planned the evaluation presented in this chapter. Still, the fact that they are near-identical is fortuitous. We can cleanly identify the additional effect of a meta-analysis compared to an individual study, holding constant the substantive conclusions furnished by the evidence. Few existing sets of studies meet these criteria.

We integrated a field experiment into a one-day "evidence summit" held in Washington, D.C. Our research subjects were mostly mid-level and senior policymakers and practitioners active in the capital. They included both federal government employees, as well as individuals working at think tanks and not-for-profit organizations. The sample, therefore, comprises exactly the types of people that policy-focused political scientists seek to reach and persuade.

We randomly varied the order in which event attendees were exposed to different presentations of the Metaketa results, in addition to a placebo condition and a non-Metaketa study on the same topic-area. By taking frequent measurements of outcomes, the design allows us to assess whether access to a larger dose of coordinated studies affects beliefs and out-of-sample prediction accuracy. It also sheds light on a variety of other hypotheses about cumulative learning, which we now develop.

[2] Weiss et al. (2008): 33.

12.2 EXPECTATIONS

12.2.1 Existing Evidence

Before describing our experiment, we review the state of knowledge on three questions: Is external validity possible in the social sciences? Are policymakers responsive to evidence? And if so, what role might additional evidence play in shaping beliefs about the effects of a policy or program?

For our experiment to generate meaningful results, external validity must be attainable. In common usage, studies are said to lack external validity when the manner or environment in which they were performed is unrepresentative of a larger set of cases. But simple statistical theory suggests that the problem may run deeper. To make inferences about a population of interest, we typically investigate a sample of cases drawn randomly from that population.[3] Yet studies designed to estimate causal effects are rarely – if ever – chosen in this manner. Without random sampling, generalizing from one context to another may be a tenuous exercise. This is especially true for interventions whose effects are heterogeneous – moderated heavily by external factors, or changeable over time.[4]

A sizable body of work has sought to measure whether external validity obtains. On balance, the empirical record is quite optimistic. There are a number of cases of successful extrapolations from studies to broader populations. For example, DellaVigna and Pope (2017) show that academic experts guess the results of online behavioral experiments with remarkable accuracy. Tetlock and Gardner (2016) identify "superforecasters" who consistently predict national security and foreign policy outcomes on the basis of very limited information. Strikingly, in betting markets, both experts and nonexperts are able to forecast at high rates whether experimental studies in psychology and economics will replicate (Dreber et al., 2015; Camerer et al., 2016).[5]

[3] Since the expected value of this sample average is known to equal the population average, extrapolating outside of the sample is trivial in this case.

[4] Thinking formally, prediction could still be possible, even in the absence of randomly sampled studies if, for example, there are informative beliefs about the sampling process.

[5] To be sure, some interventions have been shown to succeed in one place only to fail when tried elsewhere (Grossman, Humphreys, and Sacramone-Lutz, 2019). For further examples on both sides of this debate, see Banerjee and Duflo (2009): 160–61.

Our experiment provides a further test of the viability of out-of-sample prediction, and thus of external validity writ large. It diverges from existing literature, however, which mostly elicits forecasts about the results of laboratory studies. By contrast, we measure and verify predictions about a corpus of field experiments conducted in five different countries worldwide. Ex ante, this represents a much harder test.

Next, our experiment speaks to the issue of whether policymakers internalize the lessons from rigorous evaluations. The literature highlights several pathologies. Policymakers may process research in "directional" (i.e., biased) ways. Individuals frequently discount evidence that contradicts prior judgments, and over-weight confirming evidence (Molden and Higgins, 2012; Nyhan and Reifler, 2015; Baekgaard et al., 2017). Institutional constraints might also pose a hurdle in this regard. NGOs or government agencies that have sunk significant resources or political capital into backing particular interventions may be unwilling to change course, no matter what the evidence says. Similarly, evidence might be sought out, but only selectively – to legitimate decisions that have already been taken (Patton, 2015). There is a danger that policymakers fixate on studies demonstrating large, significant results, without paying attention to replicability or context specificity. It is possible, too, that the skills needed to properly decipher and interpret statistical evidence are in short supply. Policymakers, mindful of questionable research practices, might reasonably distrust much of the evidence being produced. In short, policymakers could well predict how evaluations will play out, but still not have that information affect their beliefs and decision-making.

While these arguments are plausible, hard evidence about elite behavior is scant. Most of what we know about the policymaking process comes from qualitative research. Our experiment enables us to test systematically whether a relevant sample of active policymakers suffer from consistency bias, recency bias, and the overweighting of statistically significant research findings. We can also measure the magnitude of updating over different types of beliefs.

The forgoing discussion addresses whether policymakers are responsive to evidence at all. A final question is whether exposure to additional evidence on a given topic will affect how policymakers update their beliefs. In a recent paper, Vivalt and Coville (2017) propose a simple learning model in which policymakers have to decide between two programs: one whose impact is certain, and another whose impact is uncertain. Policymakers seek to maximize program effects. A research study provides a signal about the effectiveness of the uncertain program, and policymakers update beliefs in response to the information. The marginal benefit of carrying out an extra study is the probability that its evidence will be pivotal in determining which program the policymaker selects, multiplied by the anticipated gains from choosing that program instead of the other one. Simulations, as well as real data on policymakers' priors, suggest that studies are rarely pivotal in practice. Studies matter more for updating when there is less evidence available, and when there is greater initial uncertainty about a program's impact.[6]

While important, these results only employ data on policymakers' priors; actual updating in response to new information is not observed. Moreover, an assumption behind the model is that policymakers are rational Bayesians – something that may or may not be true. Our study can help plug both these evidentiary gaps.

12.2.2 Hypotheses

Prior to implementing the experiment we registered six hypotheses about how the Metaketa evidence might influence policymakers. Our hypotheses focused on the effects of learning about (a) an individual Metaketa study; (b) a meta-analysis of the Metaketa studies; and (c) an external study, conducted outside the Metaketa project but focused on the same questions and employing a credible research design. We purposively excluded one single study from the meta-analysis. Predictions were then made about this omitted "unseen" study.

Our first two hypotheses relate to the changes in predictions we expected over time owing to exposure to all types of evidence:

- H1: Subjects will be more accurate in forecasting the results of an unseen study at endline compared to baseline, having been exposed to all types of presentations (a nonexperimental comparison).
- H2: Subjects will be more confident in their forecasts regarding the results of an unseen study at endline compared to baseline (a nonexperimental comparison).

[6] Beynon et al. (2012) conduct an experiment on the impact of policy briefs, a commonly used tool for disseminating evidence to policymakers. Overall, the study indicates that policy briefs summarizing research in a simplified format do little to shift perceptions about what kinds of interventions work, although policymakers with flat priors do update somewhat.

Three hypotheses relate to the separate effects of these three sources of information:

- H3: Subjects exposed to an individual study will be more likely to make a correct prediction about the result of the unseen study than subjects exposed to the placebo information (an experimental comparison).
- H4: Subjects exposed to the external study will be more likely to make a correct prediction about the result of the unseen study than subjects exposed to the placebo (an experimental comparison).
- H5: Subjects exposed to the (leave-one-out) meta-analysis will be more likely to make a correct prediction about the result of the unseen study than subjects exposed to the placebo, or to an individual Metaketa study, or to the external study (experimental comparisons).

The final hypothesis centers on decision-making about the use of programming funds:

- H6: Subjects exposed to the meta-analysis will reduce the percentage of funds allocated to the voter information intervention compared to subjects exposed to the placebo, or to an individual Metaketa study, or to the external study (experimental comparisons).

12.3 DESIGN

To test these hypotheses, we conducted a field experiment with policymakers and program officers based in the Washington D.C. area, and working in the fields of governance and democracy promotion. The experiment was embedded within a one-day "evidence summit," billed as an opportunity for interested parties to learn the results of the Metaketa I initiative. The event lasted approximately four hours. The setting for the experiment was naturalistic insofar as seminars, workshops, and one-day conferences occur regularly in the US capital.

On entering the event space, subjects were asked to provide informed consent and were handed a manila envelope.[7] They were asked to hold on to their assigned envelope for the rest of the day. Paper-clipped to the

[7] The experimental protocol was approved by the Institutional Review Board at the University of California, Berkeley, under case number 2017-04-9779.

front was a list of instructions and an entry survey. Subjects completed the entry survey, but were told not to open the envelope until instructed to do so.

Participants then gathered in the main auditorium. Here, we introduced the EGAP network and the Metaketa research model. We went on to provide key background information on Metaketa I: its common research question, the intervention, operational definitions of "good" and "bad" news, and the harmonized outcome measures. Next, each of the six Metaketa I teams that had completed their studies delivered a short (five minute) presentation describing the country in which their experiment was fielded, the type of information delivered to citizens, its mode of delivery, the politicians about whom information had been provided, and the study's sample size. Importantly, none of the presentations mentioned any study results.

After the teams had finished speaking, we briefed the audience on the upcoming experimental part of the event. We told subjects that they could now open their envelopes, which contained three items: results-free summaries of each Metaketa study; a personalized itinerary telling the participant which room to go and when; and five identical outcome sheets stapled together in a bundle. All five sheets elicited predictions and beliefs about one of the six Metaketa studies. (We refer to this as the subject's "unseen" study, since they did not get to learn its actual results until after the experiment had concluded.)

The logistical directions were straightforward. Subjects were to go to the four rooms indicated on their personalized schedule at the prescribed times, and to complete an outcome sheet immediately after hearing each presentation. To avoid spillovers, we emphasized the importance of not conferring with other attendees while the experiment was in progress. We said that participants should fill in the top-most outcome sheet in their packets before leaving the auditorium, providing us with pretreatment predictions and beliefs.

Once these preliminaries were over, participants moved to a nearby block of classrooms. The classrooms were numbered one through eight, corresponding to the room numbers given on the participants' schedules.

12.3.1 Treatments

The experiment consisted of four rounds. Subjects were exposed to four types of presentations – one presentation per round – in an order that varied randomly by individual.

The four presentation types were as follows:

1. **Single (Individual Metaketa study).** Participants were exposed to the results of one of the six completed Metaketa I studies.
2. **Meta (Leave-one-out meta-analysis).** In preparation for the event, we generated all six possible versions of the Metaketa meta-analysis that had one of the six studies omitted. Participants were exposed only to the version of the meta-analysis that did not include their unseen study.
3. **Ext (External study).** We presented every participant with the results of a non-Metaketa study: Ferraz and Finan's "Exposing Corrupt Politicians: The Effect of Brazil's Publicly Released Audits on Electoral Outcomes."[8] This influential paper, like Metaketa I, addresses how the availability of information impacts electoral behavior. Unlike the Metaketa studies, however, Ferraz and Finan report large, statistically significant effects. In particular, they find that the release of corruption audits in Brazil significantly depressed the vote share of incumbent mayors, notably in places where the audits unearthed a greater than median number of infractions (akin to the Metaketa's "bad news" condition). Unlike the Metaketa studies, they do not investigate the effects of information on voter turnout.
4. **Placebo (Placebo condition).** This consisted of a presentation about the forthcoming Metaketa II, III, and IV initiatives. It was uninformative about the results of Metaketa I.

In total, there were fourteen unique presentations: six individual studies, six meta-analyses, the external study, and the placebo. Because multiple presenters were involved, it was important to standardize the presentations as much as possible. All results derived from parsimonious analyses without covariate adjustment.[9]

[8] Ferraz and Finan (2008).

[9] However, fixed effects for blocks used in the randomization were included in the models. The presentation slides are given in the online appendix (as examples) and in the replication materials (in full). The meta-analysis presentations were based on an early version of the meta-analysis which contained a coding error for the Burkina Faso study. Specifically, the codings of good and bad news were mistakenly reversed in the presented findings that involved this study. However, it is important to highlight that the estimated effects remain null – both in the meta-analysis and in the Burkina Faso study results – before and after we fixed this mistake.

They were presented as simple bar plots showing average outcomes in treatment and control groups. To make style and structure as uniform as possible, presenters rehearsed their talks jointly in advance of the event. We allowed respondents to ask clarifying questions at the end of each presentation. However, they were not permitted to discuss comparisons across the studies.

12.3.2 Outcomes

Over the course of the experiment – at baseline, and at the conclusion of each round – participants filled in five outcome sheets.[10] These were deposited in bins as participants exited each room. The sheets contained six questions.

For the first set of questions, participants were asked to guess what was the estimated effect of the bad news information on turnout and vote choice in the unseen study they had been assigned. They could circle one of three options: positive and statistically significant; negative and statistically significant; or no statistically significant effect.[11] Two additional questions instructed participants to state, using a three-point scale, how confident they felt about the predictions they had given.

The sheet's remaining questions probed beliefs about the effectiveness of the intervention. Focusing still on the unseen study, one question asked whether participants thought that receiving the bad news information would have made voters more or less likely to vote for the incumbent politician or party, regardless of whether they expected the study to have found statistically significant effects.

The final question sought to capture broader beliefs about the intervention's worth. We asked participants to imagine that an organization had put them in charge of a program to improve political accountability in developing democracies. They had $1 million to spend on three

[10] A unique subject ID tied outcome sheets to participants. Participants were required to fill out the same prediction sheet five times. Participants were instructed that they were free to change their answers or keep them the same across rounds, as they saw fit.

[11] To help participants, the sheet included a short primer on statistical significance. It was explained as follows: "The difference between measured outcomes in treatment and control groups is said to be statistically significant when such a difference is highly unlikely to be due to random chance alone. On the other hand, a difference is said to be not statistically significant when it is quite possible that it is due to chance." We also clarified the meaning of a "positive" and "negative" effect in relation to the specific outcomes.

ongoing initiatives. Their job was to allocate the funds in the most cost-effective manner. The three initiatives were:

- Providing information to voters about politician performance in the run-up to an election;
- Giving special training to politicians and bureaucrats that tries to help them better perform their duties;
- Funding nonpartisan observers to monitor upcoming elections and check for irregularities.

The original outcome sheet is given in the online appendix. The six dependent variables, which are based directly on the questions given on this sheet, are then as follows:

1. **Vote (Est.).** Whether or not the respondent correctly predicted vote choice for the unseen study (based on outcome sheet, question 3).
2. **Vote (Cert.).** Respondent's confidence in the vote choice prediction (based on question 4).
3. **Turnout (Est.).** Whether or not the respondent correctly predicted turnout for the unseen study (based on question 1).
4. **Turnout (Cert.).** Respondent's confidence in the turnout prediction (based on question 2).
5. **Vote (Real).** Absolute scale-point difference, squared, between the response given, and the "null" option on the scale – i.e., "Neither more nor less likely" (based on question 5).[12]
6. **Allocation.** Proportion of funds allocated to the voter information intervention (based on question 6).

12.3.3 Randomization

The basic idea behind our experimental set up is to expose each subject to all four types of presentation – Single, Meta, External, and Placebo – in a randomized order, and to randomly assign each subject an "unseen" study, about which they declare five sets of predictions over the course of the event. Note that there are six versions of the leave-one-out meta-analysis presentation and six single study presentations. For the design,

[12] Consider, for example, a respondent who answered "Much more likely" to question 5. Their outcome value would be recorded as 4, since the answer they provided is two scale points away from "Neither more nor less likely," and $2^2 = 4$. Meanwhile, a respondent who answered "Somewhat less likely" would have an outcome value of 1.

the specific meta-analysis and single study to be viewed is randomly chosen for each participant. This ensures that our findings do not pick up beliefs about – or idiosyncratic features of – any one particular study or context. We stipulate a given subject's unseen study to be the one excluded from his or her assigned meta-analysis (e.g., for a subject assigned the meta-analysis without Benin, her unseen study is Benin). The randomization is constrained so that an individual's assigned single study cannot be the same as her unseen study. A key virtue of the randomization is that it safeguards against order or period effects biasing our results.

Concretely, the randomization procedure is implemented in three steps.

First, the fourteen unique presentations are randomly assigned to rooms and time slots. This is a restricted randomization in which, for logistical reasons, the Uganda 1 and Mexico single studies are randomly assigned to different time slots in room 1. Then the other four studies are assigned randomly to rooms 2 and 3 in rounds 1 and 2. The resulting pattern of assignments to slots and rooms is repeated for rounds 3 and 4. The same pattern is then replicated for the leave-one-out meta-analysis: the six presentations are allocated to three rooms across four rounds, with each meta-analysis appearing twice. With the external study and the placebo allocated to rooms 7 and 8 for all rounds, this step results in a full timetable indicating what is presented in each of eight rooms over four rounds (see Table 12.1).

Second, conditional on the room assignments, 192 treatment profiles are identified. Each profile is a sequence of four rooms to visit across the four rounds. These profiles form the exhaustive set of profiles in which each contains one single country study, one meta-analysis, one placebo, and one external study viewing.[13]

In the final stage, an ordering of profiles is preassigned to subjects 1 through *n* (where *n* is the total number of subjects), ordered as they register. Randomization ensures that the specific treatment profiles assigned

[13] To see where 192 comes from, label the first three single studies shown A, B, and C, and the second three D, E, F. Note that a subject viewing single study A could view a meta-analysis excluding study B and C in one of two orders (i.e., in rounds 1–3 or 3–1) and could view a meta-analysis excluding study D, E, or F in four ways (i.e., 1–2, 1–4, 3–2, 3–4). Given the same possibilities for any other study there are $6 * (2 * 2 + 3 * 4) = 96$ position combinations for the single study and the leave-one-out meta-analysis to take. With the positions of these two studies fixed, the placebo could either precede or follow the external study. This produces 96 * 2 = 192 combinations.

TABLE 12.1 *Room allocations in each round of the experiment. Country names refer to single studies. Study names preceded by "w/o" refer to meta-analyses with that one study left out.*

	Round 1	Round 2	Round 3	Round 4
Room 1	Uganda 1	Mexico	Uganda 1	Mexico
Room 2	Brazil	Uganda 2	Brazil	Uganda 2
Room 3	Burkina Faso	Benin	Burkina Faso	Benin
Room 4	Meta w/o Uganda 1	Meta w/o Mexico	Meta w/o Uganda 1	Meta w/o Mexico
Room 5	Meta w/o Brazil	Meta w/o Uganda 2	Meta w/o Brazil	Meta w/o Uganda 2
Room 6	Meta w/o Burkina Faso	Meta w/o Benin	Meta w/o Burkina Faso	Meta w/o Benin
Room 7	External Study	External Study	External Study	External Study
Room 8	Placebo	Placebo	Placebo	Placebo

are unrelated to subjects' time of registration. We use blocked random assignment of subjects (blocking across treatments) to guarantee that the number of subjects across all rooms is the same in round 1. Were there to be 192 subjects, perfect balance would be maintained in all rounds. With fewer than 192, balance is maintained in the first round but lost thereafter (although it is maintained in expectation). In practice many fewer than 192 subjects attended.

The randomization code is provided as an R script in the online appendix.

12.3.4 Sample

We employ a convenience sample of policymakers based in Washington, D.C. EGAP staff, in conjunction with EGAP's membership, compiled a list of 284 possible attendees. We targeted individuals working in international development, democracy promotion, and governance. Subjects were recruited via email, sent out three weeks before the event. 124 individuals confirmed attendance, while fifty-five invitees participated in the experiment.

Sample characteristics are summarized in Online Appendix Table H1. Most attendees were active policy practitioners. Just over half worked in not-for-profit organizations, and 21 percent were government employees.

TABLE 12.2 *Distribution of subjects' priors – as reported in the entry survey – on the effects of providing voters with bad news on turnout and voting for the incumbent*

Expectation	Turnout	Vote
Much less likely	0.02	0.07
Somewhat less likely	0.43	0.72
No difference	0.33	0.09
Somewhat more likely	0.2	0.11
Much more likely	0.02	0

Proportions in cells may not sum to 1 due to rounding.

(Six participants were academics.) 54 percent of subjects are male. 71 percent hold either mid-level or senior-level positions within their organizations. The sample was highly educated: 82 percent held either a master's degree or doctorate. Moreover, 60 percent indicated that they were either "somewhat proficient" or "very proficient" in statistics, and 70 percent reported having worked on impact or policy evaluations in the past. In short, the sample largely consisted of individuals familiar with quantitative research methods, making them well equipped to engage with the results of the Metaketa.

We also gathered data on subjects' prior beliefs about the effects of "bad news" information on voting behavior – framed in the abstract, and not in relation to the specific Metaketa studies. The distribution of these priors is shown in Table 12.2. Interestingly, most participants believe that information matters for turnout propensity, although there is disagreement on the direction of the effect, with the plurality expecting that bad news depresses turnout (consistent with the findings of Chong et al., 2015). 78 percent believe that bad news makes voters less likely to vote for incumbents, but most expected these effects to be modest. About one in twelve expected strong effects on vote choice and a small share expected positive effects (consistent perhaps with Vaishnav, 2017).

12.3.5 Estimation Strategy

We carry out two kinds of analysis: (a) causal estimates of the impact of exposure to one type of presentation versus another; (b) nonexperimental, before/after comparisons of participants' predictions and beliefs about the unseen study. For the experimental analyses, we estimate average causal effects within each pairwise combination of the four aggregated treatment categories – thus six possible pairwise comparisons in all.

Because policymakers are a hard-to-reach population, we anticipated a small sample size. To maximize statistical power, each experimental comparison uses one data point from every subject. Doing this requires a more involved statistical analysis.

For ease of exposition, suppose the four main treatment categories (i.e., presentation types) have the initials A, B, C, and D, and suppose our interest lies in comparing the effect of viewing A instead of B.

We classify participants into one of three strata, depending on the randomized schedule they were assigned:

- **Stratum 1.** Round 1 responses by participants exposed to either A or B in Round 1;
- **Stratum 2.** Round 2 responses by participants exposed to either A or B in Round 2, and who were not exposed to either A or B in Round 1;
- **Stratum 3.** Round 3 responses from participants exposed to either A or B in Round 3, and who were not exposed to either A or B in either Round 1 or Round 2.

Conditional on stratum, the probability of being in condition A or B is 50 percent, and so simple fixed effects analysis is sufficient to take account of differences across strata. The estimating equation is as follows:

$$Y_i = \kappa + \delta Treatment_i + \theta_s + \epsilon_i \qquad (12.1)$$

where κ is a constant, and θ are the fixed effects for strata s. We want to know δ, the estimated average treatment effect associated with outcome Y of exposure to A rather than B.

For the nonexperimental analyses, we restrict the estimation sample to responses provided in the baseline and endline surveys only, yielding two observations per participant. We then run first-differences OLS regressions of the following form:

$$Y_{i,t} = \alpha + \beta Endline_i + \gamma_i + \epsilon_{i,t} \qquad (12.2)$$

Here, $Y_{i,t}$ stands in for one of our six dependent variables; i indexes participants; t indexes time, where $t = \{baseline, endline\}$; γ_i are subject fixed effects; α is a constant; and $\epsilon_{i,t}$ is the error term. The parameter of interest is β: the within-subject estimated difference for a given Y across outcomes measured in the baseline and endline surveys.

12.4 RESULTS

We now discuss the results of the experiment, which we estimate following the specifications in equation (12.2).

12.4.1 Effects of Exposure to Presentation Types

The study's main findings are plotted in Figure 12.1. The paired comparisons being made in each row are listed on the vertical axis. The dot and whiskers plots visualize the estimated average treatment effects and their associated confidence intervals. Each column summarizes effect estimates for one of our six dependent variables.

The first row of Figure 12.1 presents the strongest experimental contrast: the differences in responses for subjects exposed to the meta-analysis (of five out of six Metaketa studies, with the subject's unseen study left out) compared to the external study. Recall that all versions of the meta-analysis report null effects, whereas the external study shows negative, statistically significant effects of bad news on vote choice. Given these divergent findings, we expect the gap in our experimental outcomes – in terms of predictions and beliefs about the intervention – to be widest when juxtaposing these two groups.

This is what we find. Looking at the prediction tasks, Figure 12.1, row 1 demonstrates large differences in the expected direction for both vote choice (column 1) and turnout (column 3). Note that these are dichotomous outcome measures taking one when the respondent's prediction aligns with the result reported in the unseen study, and zero otherwise. Viewing the meta-analysis instead of the external study significantly increases the likelihood that respondents correctly forecast the results of the unseen study.

The effects on confidence in these predictions (Figure 12.1, row 1, columns 2 and 4) are less pronounced and not statistically significant. This makes some sense. In principle, both the external study and the meta-analysis could increase participants' certainty about their predictions relative to baseline. Still, the coefficients are positively signed, suggesting that the meta-analysis produces greater gains in certainty than exposure to the external study. Since the meta-analysis

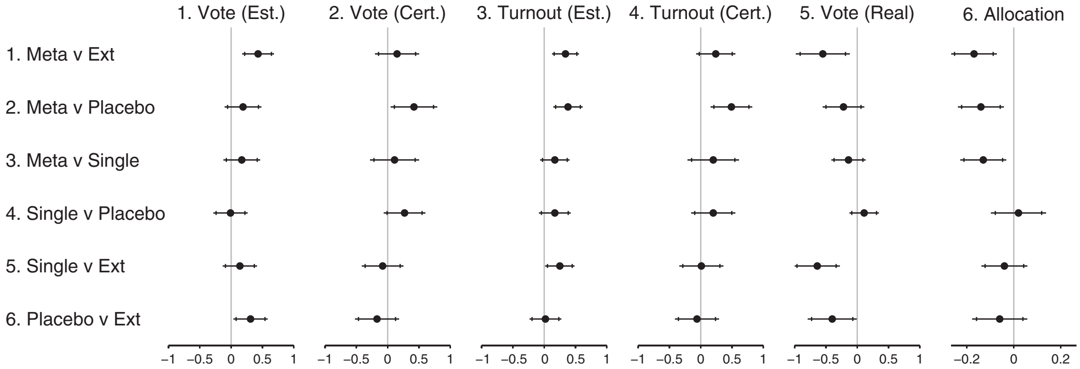

FIGURE 12.1 Causal effects of exposure to presentation types, and 95 percent confidence intervals (ticks denote 90 percent confidence intervals). The experimental comparisons being made are shown by row. For each paired comparison, the first listed item is the treatment condition, and the second item is the comparison condition. The six dependent variables, listed by column, are as follows: (col. 1) whether or not respondent correctly predicted vote choice in the unseen study; (col. 2) respondent's confidence in vote choice prediction; (col. 3) whether or not respondent correctly predicted turnout in the unseen study; (col. 4) respondent's confidence in turnout prediction; (col. 5) beliefs about intervention's "real" effectiveness; (col. 6) proportion of funds allocated to the voter information intervention. *N* is 46–49 for each analysis.

was the more informative about the unseen study, this chimes with expectations.

We also see sizeable effects in Figure 12.1, row 1 for beliefs about the intervention's effectiveness (column 5). Interpreting the column 5 coefficients requires some care. Responses to the original survey question were given on a five-point ordinal scale, asking how likely it is that the intervention affected vote choice. For the purposes of the analysis, responses are re-coded so as to create a measure of the absolute scale-point difference, squared, from what the Metaketa suggests is the best answer to this question (i.e., no effect). A negatively signed coefficient thus indicates a shrinking of this distance, with respondents moving towards what the six studies point to as the intervention's impact; a positive value indicates that participants' beliefs are moving away from the conclusion of the studies. In row 1, column 5, we see that exposure to the meta-analysis induced significant movement towards the Metaketa's null result, when compared to those who viewed the external study.

There are very substantial differences in Figure 12.1, row 1 on decisions about how to allocate expenditures (column 6). The exercise was to divide a fixed amount of money ($1 million) between three democracy promotion schemes. The dependent variable used in column 6 is the fraction of total funds allocated to "Providing information to voters about politician performance in the run up to an election," in other words, the Metaketa intervention. The statistically significant negative effect on allocation decisions in row 1, column 6 corresponds to a 17 percent drop in allocations when the meta-analysis is seen, as opposed to the external study. Policymakers' willingness to transfer funds away from the intervention is reassuring in view of the strong meta-analysis evidence that providing information to voters is ineffectual.

It is important to highlight that the effects displayed in Figure 12.1, row 1 are relative effects. They could be driven by the negative impacts of the meta-analysis on beliefs about the intervention's effectiveness, the positive impacts of the external study, or both simultaneously. By looking at a larger set of experimental comparisons, we can unpack which of these effects predominates.

The other analyses displayed in Figure 12.1 help to decompose these results. We next compare the effects of the meta-analysis against the placebo, and the placebo against the external study. Focusing on the predictions about vote choice (column 1), we observe that the effects of the meta-analysis and the external study push in opposite directions. Exposure to the meta-analysis boosts predictive accuracy relative to the placebo (row 2, column 1). Exposure to the placebo boosts accuracy relative to the external study (row 6, column 1), which of course indicates that the external study is leading respondents to mispredict the results of the unseen study at higher rates. Interestingly, the effects of the external study on the vote choice prediction in row 6, column 1 turn out to be larger in absolute magnitude than the effects of the meta-analysis (row 2). Put simply, the external study moved predictions more than the meta-analysis.

The picture is reversed for the turnout predictions in Figure 12.1, column 3. The estimates here suggest that the difference in prediction accuracy between groups exposed to the meta-analysis and the external study (in row 1) are attributable wholly to the effects of the meta-analysis. (To see this, contrast the significant, positive effect in row 2, column 3, with the null effect in row 6, column 3.) Since the external study did not analyze turnout at all, the noneffect in row 6, column 3 is both unsurprising and logical.

The differences in effects on beliefs about intervention effectiveness and allocations are split between effects driven by the meta-analysis and effects produced by the external study. Again, we use the placebo as a benchmark. The meta-analysis gives rise to a bigger change in allocation decisions than the external study (contrast Figure 12.1, rows 2 and 6 in column 6). But the external study more strongly alters beliefs about the "real" effects of the intervention (row 2 versus row 6 in column 5).

What of the differences in the effects of single Metaketa studies compared to the meta-analysis? This is an important comparison for validating the Metaketa approach. The initiative is built on the notion that multiple, coordinated field experiments provide a stronger basis for assessing a program's impact than individual studies performed and published in isolation. Our hope, therefore, is that policymakers update more after viewing a meta-analysis than an individual Metaketa study.

The data lend only partial support. Two sets of comparisons in Figure 12.1 are illuminating. First, scanning across columns, it is noteworthy that the effect of the meta-analysis compared to the placebo (row 2, significant for four of six outcomes) is much stronger than the effect of single studies compared to the placebo (row 4, null for all outcomes). On this barometer, subjects appear to assign substantially more weight to the meta-analysis.

However, the second, more exacting test is the head-to-head comparison between the meta-analysis and the single study presented in Figure 12.1, row 3. Qualitatively, the effects for outcomes in this row are positive in columns 1–4, and negative in columns 5–6. This is as we hypothesized; it is consistent with respondents placing extra weight on the meta-analysis over and above the single study. But in only one case does a coefficient in row 3 rise to the level of statistical significance – the one for allocation decisions in column 6.

Perhaps surprisingly, the differences in effects of the single Metaketa study versus the external study are large in two cases: the turnout prediction (Figure 12.1, row 5, column 3) and beliefs about program effectiveness (row 5, column 5). The direction of these two effects shows that participants are putting more credence in the single Metaketa study – appropriately so, given the task at hand.[14]

In general, the causal results show considerable responsiveness to information. They reveal sophistication in extracting information from distal studies, with greater weight placed on more relevant studies. Encouragingly, we do not see evidence that policymakers are captivated by studies that demonstrate large, significant effects. Perhaps the most disappointing result from the Metaketa point of view is that the differences in effects of single Metaketa studies versus multiple studies – though substantively large – are not significant in all specifications.

12.4.2 Baseline/Endline Comparisons (Nonexperimental)

The experimental analyses just discussed examine the effects of the different treatments relative to one another. By the end of the final round of the experiment, all participants had seen all treatments. We now explore the net impact of exposure to the full collection of presentations.

Figure 12.2 presents the average within-subject differences in outcomes across baseline and endline surveys. We urge caution in attaching a causal interpretation to these over-time estimates. Period effects – for example, growing fatigue or frustration on the part of subjects – could

[14] There is a proviso regarding the analyses in Figure 12.1. Two of the experimental contrasts in Figure 12.1 – Single versus Placebo (row 4) and Single versus Ext (row 5) – employ observations from subjects already exposed to the meta-analysis. In similar fashion, another two contrasts – Meta versus Placebo (row 2) and Meta versus Ext (row 1) – employ outcomes from subjects already exposed to a single Metaketa study. This unavoidable feature of the randomization is not a source of bias, but it may lead to attenuated treatment effects. In Figure H1 in the online appendix, we rerun the analysis, excluding subjects already exposed to these pieces of information. The effects appear somewhat stronger, as expected.

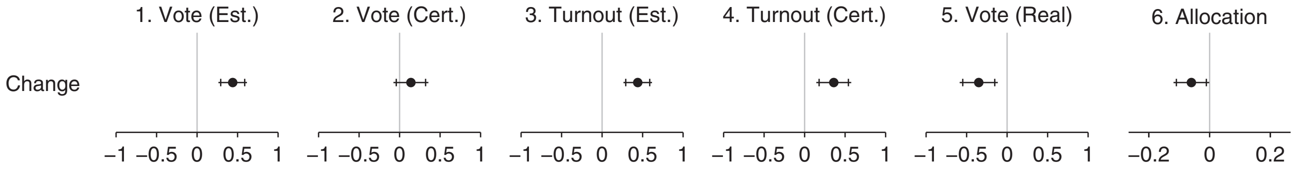

FIGURE 12.2 Differences in outcomes before and after exposure to the full collection of presentations. Dependent variables are as listed in Figure 12.1. *N* is approximately 96, with two observations per subject for each outcome.

conceivably drive patterns seen in the data.[15] That said, the experiment lasted only eighty minutes, making the inter-temporal changes more compelling.

We start by considering the results for study predictions. In Figure 12.2, we see a large uptick in predictive accuracy for both vote choice (column 1) and turnout (column 3). At baseline, 27 percent of respondents believed that the bad news treatment would have no statistically significant effect on vote choice; by endline, that figure had risen to 71 percent. The corresponding numbers for turnout are 44 percent and 88 percent, respectively. Viewing the four presentations in their entirety was associated with improved ability to make out-of-sample predictions.

Did participants' confidence in these forecasts go up as the experiment proceeded? Only modestly so, according to the self-reports (Figure 12.2, columns 2 and 4). Confidence was already low at baseline: on a three-point scale ranging from zero (low) to two (high), average confidence stood at just 0.97 (for vote choice) and 0.83 (for turnout) prior to the presentation of study results. By endline, these averages had moved up by 0.14 and 0.36 scale-points, respectively. Only the confidence increase for the turnout prediction is statistically significant. Subjects appear to be cognizant of difficulties in guessing how the results of any particular study might turn out.

Does participation in the event shape beliefs about the "real" impact of the intervention? Column 5 of Figure 12.2 suggests that participants were somewhat persuaded that the bad news information would have negligible impacts on voter behavior: the distance-squared index drops by a statistically significant amount. In a directional sense, therefore, respondents' expectations about the real-world impact of the intervention track how they are updating over the study results, with both sets of beliefs converging on the null.

A different perspective on this result can be gleaned by inspecting the raw, untransformed responses to the survey question used to construct the column 5 dependent variable. Strikingly, most respondents at endline – 61 percent – continued to report that voters presented with the bad

[15] For instance, one could imagine that fatigue or frustration might have led participants to offer an increasing number of neutral responses to the survey questions as the event progressed. This would be problematic in our application, since what we code as being the "correct" response to questions is usually the neutral one.

news information would be "Somewhat less likely" to vote for the incumbent politician. Only 32 percent said that the information would make voters "Neither more nor less likely" to back the incumbent – although this represented a large increase relative to the 10 percent who gave that answer at baseline. Learning about the Metaketa results swayed some participants at the event. But it was not enough to convince the majority that the intervention was ineffective.

Finally, we examine respondents' decisions about how to divide funds (Figure 12.2, column 6). The temporal change in this variable, although negative as expected, is arguably modest in size. There is a six percentage point reduction in the share of funds assigned to voter information programming at endline compared to baseline – a marginally significant difference. This substantively minor shift is perhaps surprising in view of the volume of Metaketa evidence showing the intervention to be inconsequential for voter behavior.

Why might this be the case? It could be that policymakers held fast to the idea that the intervention might be effective under certain conditions – in contexts where Metaketa studies had not been conducted. Another explanation is that policymakers thought that tweaks to the intervention itself might enhance its effects. More worrying is the possibility that policymakers, after hearing a slew of null results, negatively updated about the ability of studies to detect actually existing effects. We have some empirical traction on this issue. Respondents were asked in both the entry survey and at endline, "How valuable do you think randomized controlled trials are for identifying a program's effects in the real world?" At the start of the event, the mean response-score to this question – recorded on a five-point value scale, with 5 denoting "very valuable" – was 3.85; by endline, this figure had barely changed, standing at 3.70. Fortunately, therefore, exposure to the Metaketa null results does not seem to have heightened misgivings about the value of experimental evidence.

To sum up, we see quite dramatic increases in predictive accuracy as the experiment progressed. Beliefs about the potency of the intervention also moved, but to a lesser extent, with hypothetical allocations to the voter information program dropping only slightly. Although more participants thought that the intervention's true impact in the unseen study was essentially zero, most continued to believe at endline that the intervention would have an effect in the unseen studies, despite the raft of Metaketa evidence to the contrary.

12.5 ARE POLICYMAKERS BAYESIANS?

We see strong evidence for updating in the expected directions. But is this updating "optimal"? The standard approach to rational updating is to use Bayes' rule. One starts with a prior belief about a hypothesis and then updates in a specific way, given new data.[16]

Without a strong handle on individuals' beliefs about the probability of data under the hypothesis – for example, here the probability that study *B* would find effects if the excluded study *A* were to exhibit effects – and a good understanding of priors, assessing conformity with Bayes' rule is difficult. Nevertheless, one testable implication can be exploited. According to Bayes' rule, the order in which data is presented should not affect conclusions. A subject seeing the meta-analysis first and the results from Ferraz and Finan (2008) later should ultimately form the same view as a subject seeing Ferraz and Finan first and the meta-analysis later. We can check this prediction against patterns in the data. Specifically, we can examine the differences across treatment groups based on what information subjects received first (differences that should be large) compared to the differences across groups defined by what they received last (differences that should be negligible). In simple terms, by the end of the experiment, everyone had been exposed to the same information. If Bayesian updating is in force, therefore, all participants should converge on (roughly) the same posterior.

Table 12.3 reports empirical tests. We regress outcomes about beliefs in "real" effects (columns 1–3) and fund allocations (columns 4–6) on indicators for exposure to various presentation types. The excluded category is the external study condition. Columns 1 and 4 show the effects of round 1 treatment assignment on round 1 outcomes; columns 2 and 5 show the effects of round 4 treatment assignment on round 4 outcomes; and columns 3 and 6 show the effects of round 1 treatment assignment on round 4 outcomes.

If policymakers suffer from recency bias – heeding only the most recent information they came into contact with – then we should observe significant effects in columns 2 and 5 of Table 12.3; put otherwise, participants' round 4 (endline) responses should be influenced heavily by

[16] Specifically let $p(H)$ denote the prior about hypothesis H, let the belief about the likelihood of seeing the data that is seen under the hypothesis be given by $p(D|H)$, and the probability of seeing that data under any hypothesis by $p(D)$. One then updates using the rule: $p(H|D) = p(D|H)p(H)/p(D)$.

TABLE 12.3 *Treatment effects in individual rounds.* T *refers to treatment assignment for a given round, while* R *refers to outcomes in that round. Thus "T1 on R4" is the estimated effect of round 1 treatment assignment on round 4 outcomes.*

	Vote (Real)			Allocation		
	T1 on R1	T4 on R4	T1 on R4	T1 on R1	T4 on R4	T1 on R4
	(1)	(2)	(3)	(4)	(5)	(6)
Meta analysis	−0.879***	−0.194	−0.164	−0.195***	0.008	−0.100
	(0.295)	(0.202)	(0.204)	(0.067)	(0.093)	(0.083)
Placebo	−0.645**	0.159	−0.300	−0.112	−0.085	−0.149*
	(0.309)	(0.191)	(0.209)	(0.070)	(0.088)	(0.085)
Single study	−0.795**	−0.179	0.018	−0.087	−0.108	−0.099
	(0.295)	(0.180)	(0.204)	(0.067)	(0.083)	(0.083)
Constant	1.545***	0.750***	0.800***	0.445***	0.367***	0.390***
	(0.213)	(0.132)	(0.148)	(0.048)	(0.061)	(0.060)
Observations	45	46	42	45	46	42
Adjusted R^2	0.149	0.031	0.005	0.113	−0.006	0.006

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

their treatment assignment in round 4. Meanwhile, if consistency bias is at work, then effects should be concentrated in columns 3 and 6: respondents' answers at endline should be decided first and foremost by the information they viewed at the very start of the experiment. If respondents are Bayesians, however, we should see effects in columns 1 and 4 only: assignment should shape beliefs in round 1, but by the conclusion of round 4 – at which point, all subjects have received the same bundle of information – there should be no discernable difference according to round 4 treatment assignment.

It is the Bayesian hypothesis that receives most support in Table 12.3. The round 1 treatment effects account for 15 percent of the variation in the effectiveness outcome in column 1, and 11 percent of the variation in the allocations outcome in column 4. Other columns, meanwhile, demonstrate no evidence of either recency bias or consistency bias. We show the results for the remaining dependent variables in Table H2 in the online appendix.

12.6 CONCLUSION

In recent years, social and political sciences have undergone both a "micro-revolution" and a "credibility revolution" (Laitin and Reich, 2017, Angrist and Pischke, 2010). Advances in method and substance have dramatically expanded the scope for research findings to have tangible impacts. A central goal of the Metaketa initiative is to generate evidence that incites policy change. Yet its success in this endeavor hinges critically on how policymakers respond to coordinated research – a question on which existing literature is mostly silent.

Using a field experiment, we show that multiple, coordinated studies aid learning in a sample of active policy practitioners. Exposure to a meta-analysis (of five of the six Metaketa studies) improves out-of-sample prediction accuracy. It also changes beliefs about the effects of interventions, as well as stated preferences over resource allocation. The effects of the meta-analysis are strongest when the comparison group is presented with an "external," non-Metaketa study – one that shows results contradicting the Metaketa's. We also see effects of the meta-analysis relative to a placebo. Both of these findings help vindicate the Metaketa model. Somewhat tempering them, however, is the fact that inferences from the meta-analysis are not significantly stronger than inferences from a single Metaketa study, in most models.

Impressively, the kinds of biases often associated with learning are not visible here. Participants do not place additional weight on "positive," non-null results. Also, they do not appear to suffer from recency or consistency bias, emerging instead as rational Bayesians.

Overall the results speak to the ability of coordinated research to inform policy, and to influence out-of-sample beliefs in an "evidence-accurate" manner. More broadly, the results provide evidence of external validity, despite the fact that the set of studies implemented was in no sense a random sample from a larger population.

We began this chapter by underscoring an irony: proponents of rigorous impact evaluation have failed to demonstrate its value as a tool for affecting policy. We are open to a similar sort of critique. Our finding of external validity may itself not generalize. This could be true in two ways. First, our case is an easy one for learning. The results from the Metaketa are uniformly null. Thus the question at stake was simply how quickly participants would come to realize that the information interventions were ineffective wherever they were attempted. The inferential challenge is likely to be much more acute in the presence of heterogeneity (Vivalt, 2016). Second, the results derive from a convenience sample – albeit a highly relevant one for our application. Even still, we believe that the chapter shows at a minimum an important proof of concept:

the external validity of a study – that is, the utility of a study for making inferences to unseen cases – is not simple a problem to worry about, it is a quantity that can be empirically assessed. Policymakers can learn across studies even when these are not randomly sampled from a population, and their learning can improve both their accuracy and their confidence.