How Much Can Impact Evaluations Inform Policy Decisions?

Eva Vivalt*

Australian National University

July 30, 2017

Abstract

Impact evaluations might inform policy decisions, but the extent to which they do depends on several factors. This paper models the decision to enact a program and estimates how much a study might improve policy outcomes. We show that the marginal benefits of a study quickly fall and when a study will be the most useful in making a decision in a particular context is also when it will have the lowest external validity. Typical marginal benefits are estimated using a large set of impact evaluation results and we further explore implications using a set of real priors. The results highlight that leveraging the wisdom of the crowds can result in greater improvements in policy outcomes than running an additional study.

*E-mail: eva.vivalt@anu.edu.au.

1 Introduction

Economic studies are funded and undertaken in part to inform policy decisions, however, there are many factors that influence how much they can change policy. In this paper, we optimistically consider "ideal" policymakers who have the willingness and capacity to enact whichever program maximizes the outcomes of the program beneficiaries and who Bayesian update in response to new information. Data from twenty meta-analyses of development programs are then used to estimate the effects of an additional study and to obtain a range of parameter values to apply to the model. We find that a typical impact evaluation is unlikely to lead to large changes in policy outcomes in a given context.

Impact evaluations are sharply increasing both in number and in terms of the resources devoted to them. For example, the Millennium Challenge Corporation has committed to conduct rigorous impact evaluations for 50% of its activities, with "some form of credible evaluation of impact" for every activity (Millennium Challenge Corporation, 2009); the U.S. Agency for International Development is also increasingly invested in impact evaluations, directing 3% of program funds to evaluation[1]; behavioral insights labs conducting real-world experiments have become a part of many governments.[2] With this institutional interest, these studies seem likely to continue to rapidly grow.

At the same time, many voices, from Duflo, Glennerster and Kremer (2008) to Deaton (2010), have urged caution in interpreting results for evidence-based policymaking. Vivalt (2017a) finds considerable heterogeneity in research findings across twenty different topics in development, which *prima facie* suggests that policymakers learn relatively little from a study. This literature motivates carefully modeling how much of a policy impact we can expect from an impact evaluation. Further, there may be substantial gains from re-allocating academic talent across potential studies, and a model could help clarify which potential

[1] While most of these are less rigorous "performance evaluations", country mission leaders are supposed to identify at least one opportunity for impact evaluation for each development objective in their 3-5 year plans (USAID, 2011).

[2] *E.g.* The Social and Behavioral Sciences Team in the United States, the Behavioural Insights Team in the United Kingdom, and Behavioural Insights Units in Australia.

studies have the most promise.

We consider the case in which a policymaker faces a choice between two potential programs: a program whose performance is uncertain and an outside option which has a known effect. Policymakers update their expectations of the performance of the uncertain program as in the normal learning model. This model has been used in many other settings in which decision makers face uncertainty and update in response to new information (*e.g.* Kala, 2015; Conley and Udry, 2010), but has not been applied to this problem. An additional study could either be pivotal in a positive way, nudging policymakers to take the correct action, or pivotal in a negative way, nudging policymakers away from the optimal program. The marginal benefits of doing a study can then be calculated as the probability that the study is pivotal in determining what the policymaker would do multiplied by the expected gains from pursuing the selected option.

This paper builds on the literature in several ways. First, it shows how it can be difficult for a study to have a large effect on policy outcomes. In the model, policymakers are strictly seeking to maximize the effects of the programs they implement; they do not have other strategic considerations. Even in this charitable scenario, we find that most studies will have minor effects on policy, both because they are seldom pivotal in a policymaker's decisionmaking process and because the effects of different programs on the same outcome tend to be quite similar in our data. We are also able to estimate the effects of a study using real-world data, using real impact evaluation results as well as priors collected from policymakers, practitioners and researchers. We estimate these effects assuming either that decision-makers make decisions on their own (*i.e.* as a dictator) or else make decisions in groups using a majority voting rule and show results for both scenarios. Whether a group would make better decisions than an individual depends on the accuracy of the group's priors. Finally, the model highlights that the times when an impact evaluation will be most useful in making a decision in a particular context are also the times when it will have the lowest external validity.

The results data come from AidGrade, a non-profit research institute that collects and synthesizes data from academic studies. To date, AidGrade has conducted meta-analyses and systematic reviews of 20 different development programs.[3] Data gathered through meta-analyses provide a comprehensive view of the evidence for each topic, and data on these 20 topics were collected in the same way for each variable. Currently, the data set contains 15,024 results from 635 papers.

The unique priors data were collected largely from policymakers, researchers and practitioners attending World Bank workshops on impact evaluation. These workshops are each about one week long and can be thought of as "matchmaking" events in which teams developing projects that would like an impact evaluation as part of the project are matched with researchers who, over the course of the week, begin to design an impact evaluation with them. The people attending these events will be described in more detail later but include people who work in government agencies of various developing countries, operational staff from international organizations, and development economics researchers working closely with government agencies. These participants are not generally elected or appointed officials but more often technical advisors within a government agency, and the workshop attendees form a particularly good sample since they consist of those who are especially interested and engaged in impact evaluations.

We find that the typical impact evaluation would improve a policy decision in a given context by about 1%, or by up to 6.5% under more favorable conditions. Earlier studies, studies on interventions or outcomes with high heterogeneity in results across studies, and studies with small sampling variance have the greatest impact.

[3]Throughout, all 20 will be referred to as meta-analyses, but some did not have enough comparable outcomes for meta-analysis and became systematic reviews.

2 Theory

2.1 The Policymaker's Decision Problem

In the simplest case, a policymaker might face a choice between a program and an outside option. The program's mean effect if it were to be implemented in the policymaker's setting, θ_i, is unknown *ex ante*; the outside option's effect is θ^*. The policymaker's prior is:

$$\theta_i \sim N(\mu, \tau^2) \tag{1}$$

where μ and τ^2 are unknown hyperparameters.

The policymaker has the opportunity to observe a signal about the effect of the program by conducting an impact evaluation, as in the normal learning model. For example, this could be thought of as an evaluation of a pilot before rolling out a program. The effects are themselves are drawn from a distribution:

$$Y_{ij}|\theta_i \sim N(\theta_i, \sigma^2) \tag{2}$$

where Y_{ij} is the observed effect size of a particular study i on an individual j and σ^2 is the error variance.[4]

The impact evaluation has a cost, $c > 0$, and the policymaker needs to decide whether the value of the information provided by the signal is worth it. We assume the policymaker forms estimates of the effects of the uncertain program using Bayesian meta-analysis, described in the next section.

2.2 Bayesian Meta-Analysis

The meta-analysis literature suggests two general types of models that can be parameterized in many ways: fixed-effect models and random-effects models. Much of this exposition

[4]If n_i is the number of observations in study i, σ_i^2=σ^2/n_i is the sampling variance.

will draw from Gelman *et al.* (2013), and the interested reader is also referred to Borenstein *et al.* (2009) for a gentle introduction to meta-analysis.

Fixed-effect models assume there is one true effect of a particular program and all differences between studies can be attributed simply to sampling error. In other words:

$$Y_i = \theta + \varepsilon_i \tag{3}$$

where θ is the true effect and ε_i is the error term.

Random-effects models do not make this assumption; the true effect could potentially vary from context to context. Here,

$$Y_i = \theta_i + \varepsilon_i \tag{4}$$

where θ_i is the true effect. Random-effects models are more plausible and they are necessary if we think there are heterogeneous treatment effects ($\tau^2 > 0$). Random-effects models can also be modified by the addition of explanatory variables, at which point they are called mixed models. Both random-effects models and mixed models will be considered in this paper, however, to build intuition we will focus the exposition on the random-effects case.

2.3 Estimating a Random-Effects Model

Bayes' rule says that the posterior probability is proportional to the likelihood of the data given certain parameter values multiplied by the prior probability of those parameters. In this section we will describe the likelihood function as well as how empirically estimating this model will proceed.

Equation 1 provides the prior for θ_i, where μ and τ are unknown hyperparameters that will need to be estimated. For the likelihood, consider Equation 2; we do not have individual-level data, but can instead use sufficient statistics to form the distribution of Y_i (the study's

point estimate) given the study's true effect θ_i:

$$Y_i|\theta_i \sim N(\theta_i, \sigma_i^2) \tag{5}$$

where σ_i^2 is the sample variance.

Conditioning on the distribution of the data, we get a posterior:

$$\theta_i|\mu, \tau, Y \sim N(\widehat{\theta_i}, V_i) \tag{6}$$

where

$$\widehat{\theta_i} = \frac{\frac{Y_i}{\sigma_i^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}} \ , \ V_i = \frac{1}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}} \tag{7}$$

We still need to specify how μ and τ are found. In the basic case, we will assume a uniform prior for $\mu|\tau$ following Gelman *et al.* (2013) and update based on the data. As the Y_i are estimates of μ with variance $(\sigma_i^2 + \tau^2)$, we obtain an equation for the posterior of μ:

$$\mu|\tau, Y \sim N(\widehat{\mu}, V_\mu) \tag{8}$$

where

$$\widehat{\mu} = \frac{\sum_i \frac{Y_i}{\sigma_i^2+\tau^2}}{\sum_i \frac{1}{\sigma_i^2+\tau^2}} \ , \ V_\mu = \sum_i \frac{1}{\frac{1}{\sigma_i^2+\tau^2}} \tag{9}$$

For τ, we again use a uniform prior. We then note that $p(\tau|Y) = \frac{p(\mu,\tau|Y)}{p(\mu|\tau,Y)}$. The denominator of this equation follows from Equation 9; for the numerator, we can observe that $p(\mu, \tau|Y)$ is proportional to $p(\mu, \tau)p(Y|\mu, \tau)$, and we know the marginal distribution of $Y_i|\mu, \tau$:

$$Y_i|\mu, \tau \sim N(\mu, \sigma_i^2 + \tau^2) \tag{10}$$

This yields the posterior for the numerator:

$$p(\mu, \tau|Y) \propto p(\mu, \tau) \prod_i N(Y_i|\mu, \sigma_i^2 + \tau^2) \tag{11}$$

Putting together all the pieces in reverse order, we will construct posteriors for τ, μ and σ_i by first drawing τ, then generating $(\tau|Y)$ using τ, followed by the posterior of μ and finally the posterior of θ_i. Given the equations for the posteriors, estimating the parameters is merely a matter of making simulations, drawing from the known distributions, one step at a time. Our simulations will use 10,000 draws of τ, and after each draw we will continue to draw other values down the chain until at the end of the simulations we have 10,000 simulated values of μ and of each θ_i. We will then take the mean of the posterior distributions of τ, μ and θ_i as our estimate of these parameters.

While this exposition focused on the case of uniform priors, for robustness we will also include results that are based on a set of alternative priors elicited from policymakers, practitioners and researchers. The updating procedure is the same, simply using different priors.

2.4 Estimating a Mixed Model

We will estimate a mixed model following a similar strategy. Appendix D contains a derivation of the equations that govern the updating process. Again, we will have priors and likelihood functions for each parameter to be estimated, from which we can construct posterior distributions assuming the policymaker is Bayesian updating. To actually estimate these equations, we will again start by simulating τ, then estimating the parameters of the mixed model, then finally estimating the site-specific parameters given the estimates of τ and the other estimated parameters of the mixed model.

2.5 Solving the Policymaker's Problem and Extensions

As we saw, a random-effects Bayesian meta-analysis yields a predicted μ that depends on both the inter-study variation, τ^2, and the sampling variance, σ_i^2:

$$\widehat{\mu} = \frac{\sum_i \frac{Y_i}{\sigma_i^2+\tau^2}}{\sum_i \frac{1}{\sigma_i^2+\tau^2}} \tag{12}$$

In the simplest model, the policymaker takes their current estimate of μ as their best guess of θ_i. This would be appropriate when they do not have additional data Y_i from an impact evaluation in that setting with which to make a better estimate of θ_i. If they know θ_i is dependent on other factors and they have data on these other factors, they can make slightly more sophisticated estimates of θ_i and this can be captured using a mixed model.

This set-up enables us to calculate how often the policymaker might make the "wrong" decision and what it would cost, which would let us say for what set of conditions it might be beneficial to spend money on an additional impact evaluation. Remembering that the outside option's effects are θ^*, if $\widehat{\mu} > \theta^*$ and $\mu < \theta^*$, or vice versa, the policymaker is making a mistake costing the value of $|\mu - \theta^*|$. We can then either think of a function that assigns a cost to each $|\mu - \theta^*|$ (*e.g.* X program would have improved enrollment rates by 1 percentage point more than Y program, an improvement valued at \$Z) or keep all calculations in terms of $|\mu - \theta^*|$ (*e.g.* 1 percentage point); for simplicity, we will stick to the latter approach.

To be more precise, the marginal benefits of a study, B, are:

$$B = p^+ \cdot (|\mu - \theta^*|) - p^- \cdot (|\mu - \theta^*|) \tag{13}$$

where p^+, the probability of being pivotal in nudging policymakers towards the correct decision, is $\mathrm{P}(\widehat{\mu}_{n+1} > \theta^*, \widehat{\mu}_n < \theta^* | \mu > \theta^*)$ (n indexes the chronological order of the study) and p^-, the probability of being pivotal in nudging policymakers away from the correct decision, is $\mathrm{P}(\widehat{\mu}_{n+1} < \theta^*, \widehat{\mu}_n > \theta^*) | \mu > \theta^*)$. We consider the case that $\mu > \theta^*$, but the same

argument could be made analogously for $\mu < \theta^*$.

From this formulation, we can see several things: as $|\mu - \theta^*|$ increases, the marginal benefit B may rise, but as it continues to grow, the probability that a result will be pivotal also falls, leading to lower B; as the inter-study and sampling variance of past studies rises, the new study provides relatively more information and B rises; as the precision of the new study increases, it will affect the estimates more, and B again rises in expectation; finally, as n rises, the likelihood that the new study is pivotal falls, and consequentially B falls.

This exposition mainly concerned itself with p^+, but similar arguments would hold, in reverse, for p^-. Since we know that $p^+ > p^-$ for $\mu > \theta^*$, the net gains to B are positive. We cannot solve B analytically, as there is no analytic solution to the cumulative density function of the normal distribution, which determines p^+ and p^-. However, since we will be able to estimate the policymaker's beliefs about μ after n studies, $\widehat{\mu}_n$, we can explicitly estimate B from the data and we can also simulate it for different parameter values.

This approach focuses on the expected value of θ_i and ignores the distribution. One can imagine that in some scenarios a policymaker may care about other quantiles of θ_i. Further, one could extend the model to consider an outside option with a known distribution of possible effects. These would be straightforward extensions, but we will focus on the simplest case of a risk-neutral policymaker (or group of policymakers) facing an outside option with a known constant effect, θ^*, for whom the expected value of θ_i is all that matters.

This exercise is also limited to the costs and benefits a policymaker would face if trying to find the most efficient program with which to achieve a particular policy goal. In particular, it does not take into consideration other goals a policymaker might have, such as political concerns. While thus limited, we might think that a policymaker that seeks to make the best policy decisions in this way represents the ideal social planner.

Finally, an impact evaluation can also be thought of as providing a public good that multiple policymakers could take advantage of beyond one given context. In this case, an impact evaluation could still be worthwhile as a public good, even in some cases in which

it would not be worthwhile otherwise. Our purpose is not to critique impact evaluation but to come up with estimates of an impact evaluation's value in informing policy in a particular setting and to explore the conditions under which one would be most worthwhile. Extrapolating from these results to consider the benefits of an impact evaluation when that evaluation is used in many contexts is a straightforward exercise that depends on estimating the effect in one context as a first step.

3 Data

This paper uses two kinds of data: results data and priors data. The next two subsections will describe these data sets, in turn.

3.1 Results Data Set

This paper uses a database of impact evaluation results collected by AidGrade, a U.S. non-profit research institute founded by the author in 2012. AidGrade focuses on gathering the results of impact evaluations and analyzing the data, including through meta-analysis. Its data on impact evaluation results were collected in the course of its meta-analyses from 2012-2014 (AidGrade, 2016a).

AidGrade's meta-analyses follow the standard stages: (1) topic selection; (2) a search for relevant papers; (3) screening of papers; (4) data extraction; and (5) data analysis. In addition, it pays attention to (6) dissemination and (7) updating of results. Here, we will discuss the selection of papers (stages 1-3) and the data extraction protocol (stage 4); more detail is provided in Appendix E.

3.1.1 Selection of Papers

The interventions that were selected for meta-analysis were selected largely on the basis of there being a sufficient number of studies on that topic. Five AidGrade staff members

each independently made a preliminary list of interventions for examination; the lists were then combined and searches done for each topic to determine if there were likely to be enough impact evaluations for a meta-analysis. The list remaining after excluding topics with insufficient studies was voted on by the general public online and partially randomized. Appendix E provides further detail.

A comprehensive literature search was done using a mix of the search aggregators SciVerse, Google Scholar, and EBSCO/PubMed. The online databases of the Abdul Latif Jameel Poverty Action Lab (J-PAL), Innovations for Poverty Action (IPA), the Center for Effective Global Action (CEGA), and the International Initiative for Impact Evaluation (3ie) were also searched for completeness. Finally, the references of any existing systematic reviews or meta-analyses were collected.

Any impact evaluation which appeared to be on the intervention in question was included, barring those in developed countries.[5] Any paper that tried to consider the counterfactual of no intervention was considered an impact evaluation. Both published papers and working papers were included. The search and screening criteria were deliberately broad. The full text of the search terms and inclusion criteria for all 20 topics in this paper are available in an online appendix as detailed in Appendix A.

3.1.2 Data Extraction

The subset of the data on which we focus is based on those papers that passed all screening stages in the meta-analyses. Again, the search and screening criteria were very broad and, after passing the full text screening, the vast majority of papers that were later excluded were excluded merely because they had no outcome variables in common or did not provide sufficient data for analysis (for example, not providing data that could be used to calculate the standard error of an estimate or displaying results only graphically). The small overlap of outcome variables is a surprising and notable feature of the data. Ultimately, the data we

[5]High-income countries, according to the World Bank's classification system (2015).

draw upon for this paper consist of 15,024 results (double-coded and then reconciled by a third researcher) across 635 papers covering the 20 types of development program listed in Table 1.[6] Only 307 of these papers overlapped in outcomes with another paper on the same intervention. The small overlap of outcome variables is a surprising and notable feature of the data and suggests researchers should coordinate more.

Table 1: List of Development Programs Covered

2012	2013
Conditional cash transfers	Contract teachers
Deworming	Financial literacy training
Improved stoves	HIV education
Insecticide-treated bed nets	Irrigation
Microfinance	Micro health insurance
Safe water storage	Micronutrient supplementation
Scholarships	Mobile phone-based reminders
School meals	Performance pay
Unconditional cash transfers	Rural electrification
Water treatment	Women's empowerment programs

When considering the variation of effect sizes within a set of papers, the definition of the set is clearly critical. Two different rules were used to define outcomes: a strict rule, under which only identical outcome variables are considered alike (*e.g.* height in centimeters), and a loose rule, under which similar but distinct outcomes are grouped into clusters (*e.g.* one study may consider a subject to have anemia if their hemoglobin is less than X; another may consider a subject to have anemia if their hemoglobin is less than Y). This paper uses the strict rule wherever possible.[7]

[6]Three titles here may be misleading. "Mobile phone-based reminders" refers specifically to SMS or voice reminders for health-related outcomes. "Women's empowerment programs" required an educational component to be included in the intervention and it could not be an unrelated intervention that merely disaggregated outcomes by gender. Finally, "micronutrient supplementation" was initially too loosely defined; this was narrowed down to focus on those providing zinc to children, but the other micronutrient papers are still included in the greater data set, with a tag, and are used to examine other issues in other papers, such as publication bias.

[7]Using the loose definition preserves more data for anemia and malaria, so for these outcomes the loose definition is used.

Clearly, even under the strict rule, differences between the studies may exist, however, using two different rules allows us to isolate the potential sources of variation, and other variables were coded to capture some of this variation, such as the age of those in the sample. In total, 73 variables were coded for each paper. Additional topic-specific variables were coded for some sets of papers, such as the median and mean loan size for microfinance programs. This paper focuses on the variables held in common across the different topics. These include which method was used; if randomized, whether it was randomized by cluster; whether it was blinded; where it was (village, province, country); what kind of institution carried out the implementation; characteristics of the population; and the duration of the intervention from the baseline to the midline or endline results, among others. A full set of variables and the coding manual is available online, as detailed in Appendix A. If one were to divide the studies by all these characteristics, however, the data would usually be too sparse for analysis.

Interventions were also defined separately and coders were also asked to write a short description of the details of each program. Program names were recorded so as to identify those papers on the same program. For papers which were follow-ups, the most recent results were used for each outcome.

The data were also standardized to be able to provide a set of results more comparable with the literature and so as not to overweight those outcomes with larger scales in some analyses. The typical way to compare results across different outcomes is to use the standardized mean difference, defined as SMD $= \frac{\mu_1 - \mu_2}{\sigma_p}$, where μ_1 is the mean outcome in the treatment group, μ_2 is the mean outcome in the control group, and σ_p is the pooled standard deviation. The Appendix describes the alternative procedures used for generating the SMD when these data were not available. The signs of the results were also adjusted so that a positive effect size always represents an improvement. For the case study of the effect of CCTs on enrollment rates, raw units (percentage points) will be used.

Studies tend to report results for multiple specifications. AidGrade focused on those

results least likely to have been influenced by author choices: those with the fewest controls, apart from fixed effects. Where a study reported results using different methodologies, coders were instructed to collect the findings obtained under the authors' preferred methodology; where the preferred methodology was unclear, coders were advised to follow the internal preference ordering of prioritizing randomized controlled trials, followed by regression discontinuity designs and differences-in-differences, followed by matching, and to collect multiple sets of results when they were unclear on which to include. Where results were presented separately for multiple subgroups, coders were similarly advised to err on the side of caution and to collect both the aggregate results and results by subgroup except where the author appeared to be only including a subgroup because results were significant within that subgroup. For example, if an author reported results for children aged 8-15 and then also presented results for children aged 12-13, only the aggregate results would be recorded, but if the author presented results for children aged 8-9, 10-11, 12-13, and 14-15, all subgroups would be coded as well as the aggregate result when presented. Authors only rarely reported isolated subgroups, so this was not a major issue in practice.

A note must be made about combining data. We do not want those studies reporting results for more subgroups to have excessive weight in our analyses. Thus, where results had been reported for multiple subgroups (*e.g.* women and men), we aggregated them as in the Cochrane Handbook's Table 7.7.a. Where results were reported for multiple time periods (*e.g.* 6 months after the intervention and 12 months after the intervention), we used the most comparable time periods across papers.

Finally, one paper appeared to misreport results, suggesting implausibly low values and standard deviations for hemoglobin. This observation was excluded and the paper's corresponding author contacted.

3.1.3 Results Data Set: Descriptive Statistics

Table 2 summarizes the distribution of papers across interventions and highlights the fact that papers exhibit very little overlap in terms of outcomes studied. This is consistent with the story of researchers each wanting to publish one of the first papers on a topic. Vivalt (2017b) finds that later papers on the same intervention-outcome combination more often remain as working papers.

Table 8 in Appendix C lists the interventions and outcomes and describes their results in a bit more detail, providing the distribution of significant and insignificant results. Attention will be limited to those intervention-outcome combinations on which we have data for at least three papers.

The data have previously been analyzed for evidence of specification searching or publication bias (Vivalt 2017b), where it was found that the results from randomized controlled trials, which constitute approximately 80% of the data, exhibited few signs of bias.

The next section will describe how data on priors were collected, and then the method section will describe how we will use these data to estimate the marginal benefit of an impact evaluation.

Table 2: Descriptive Statistics: Distribution of Narrow Outcomes

Intervention	Number of outcomes	Mean papers per outcome	Max papers per outcome
Conditional cash transfers	15	18	36
Contract teachers	1	3	3
Deworming	12	13	17
Financial literacy	3	5	5
HIV/AIDS Education	5	6	10
Improved stoves	4	2	2
Insecticide-treated bed nets	1	18	18
Irrigation	2	2	2
Micro health insurance	4	2	2
Microfinance	6	4	5
Micronutrient supplementation	22	23	37
Mobile phone-based reminders	2	4	5
Performance pay	1	3	3
Rural electrification	3	3	3
Safe water storage	1	2	2
Scholarships	3	2	3
School meals	3	3	3
Unconditional cash transfers	3	10	13
Water treatment	3	8	10
Women's empowerment programs	2	2	2
Average	4.8	6.6	9.1

3.2 Priors Data

Priors were collected at 7 World Bank workshops run by the Development Impact Evaluation (DIME) research group. The workshops were conducted in Mexico City (May 2016, March 2017), Nairobi (June 2016), Lagos (May 2017), Washington, DC (May 2017, June 2017), and Lisbon (July 2017). The first two workshops were used as pilots to refine the questions and prior elicitation mechanisms. Priors were also elicited at 1 Inter-American Development Bank (IDB) workshop in Washington, DC in June, 2017.

Workshop attendees comprised policymakers, practitioners, and researchers. The workshops were each approximately one week long and were designed as "matchmaking" events

between those involved in development programs and researchers; government counterparts were paired with researchers and supposed to design a prospective impact evaluation for their program over the course of the week. Participants included program officers in government agencies of various developing countries; monitoring and evaluation specialists within government agencies; World Bank or IDB operational staff; other international organization operational staff such as technical advisors at USAID or DFID; a few staff from NGOs or private sector firms participating in a project; and academics and other researchers. Those from developing country governments are considered "policymakers"; international organization operational staff and NGO or private sector staff are considered "practitioners"; we define "researchers" to be those in academia or those who either have peer-reviewed publications or else have "research" or "impact evaluation" in their job title. In this paper, we will focus almost exclusively on policymakers and operational staff at international organizations.

Individuals were surveyed by enumerators during breaks in the workshops. Of 475 eligible attendees at the non-pilot workshops, 148 (31%) completed the survey. The main constraint was that the surveys could only be run during the typically twice-daily breaks in the workshops and during the lunch period. During the pilots, individuals were allowed to take the survey by themselves on tablets we provided and, given that many could take the survey at the same time, we had a 95% response rate. However, we changed approaches after the pilot in favor of one-on-one enumeration to reduce noise due to participants' lack of familiarity with operating the tablets and to increase attentiveness. After making this change, we still had overwhelming interest in the survey among attendees but, being limited to the breaks in the workshops, only managed to survey an average of 25 per workshop. Breaks were roughly the duration of the survey, and lunch might span 2-3 times the length of the typical break, depending on workshop timing. Thus, this response rate represents essentially the maximum number of responses that could be gathered in the allotted time, and we are confident that with additional enumerators we could have attained a substantially higher response rate. We may expect that those who managed to take the survey may have been

particularly interested in taking it or quick to approach the enumerators during a break, but we have no reason to believe that this represents a substantially different population than the universe of conference attendees. Response rates are detailed by workshop in Table 3.

In addition to gathering data at these workshops, past workshop participants were contacted by e-mail and asked to participate via video conference. The response rate was much lower in the group contacted by e-mail; of 804 eligible past workshop attendees, 59 (7%) participated in the survey. Finally, participants were also recruited at the World Bank's headquarters and at the IDB's headquarters in Washington, D.C. A table was set up by the cafeteria and passers-by were able to take the survey with a trained enumerator. 75 World Bank responses and 6 IDB responses were collected in this manner over 12.5 or 2 enumerator-days, respectively;[8] enumerators covered lunch at the IDB but full or half-days at the World Bank. Summary statistics about the various recruitment strategies and the breakdown of participants by category (policymaker, practitioner, researcher) are provided in Table 4.

Finally, a set of responses was elicited on Mechanical Turk to provide a comparison group. We required a HIT Approval Rate (%) for all Requesters' HITs greater than or equal to 95 and Number of HITs Approved greater than or equal to 50. 1,600 responses were solicited. In contrast to the policymakers, practitioners and researchers, who were interviewed one-on-one, the MTurk workers worked unsupervised.

In the survey, respondents were asked how familiar they were with several types of programs, including conditional cash transfer programs (CCTs). They were later asked for their best guess of the effect of a described CCT program on enrollment rates and asked to use slider bars to put probability weights on various effects the program might have had (see Figures 7-10 in Appendix B). Before using these sliders, participants were shown a video describing how to use the sliders and were walked through an example about predicting the weather in order to be sure that they understood the exercise. At the end of this introduc-

[8]Excluding 2 responses from support staff at each institution. These did not meet our inclusion criteria but we could not bar them from participating upfront in this context.

Table 3: Participants at Workshops

	Eligible Attendees	Surveyed	Response Rate
Mexico, May 2016 (pilot)	107	105	0.98
Kenya, June 2016 (pilot)	48	43	0.90
Mexico, March 2017	93	34	0.37
Nigeria, May 2017	75	39	0.52
Washington, DC, May 2017	44	15	0.34
Washington, DC, June 2017 (IDB)	62	10	0.16
Washington, DC, June 2017 (WB)	76	19	0.25
Portugal, July 2017	125	31	0.25
Total	475	148	0.31

This table shows the number of people surveyed at each workshop and the total number of eligible attendees. Both values restrict attention to those who could be classified as "policymakers", "practitioners" or "researchers". In addition, to be eligible to take the survey, one had to have not taken it at a previous workshop (this was primarily a concern for DIME staff) and one had to speak one of the survey languages fluently. As discussed in the text, the pilots had substantially higher response rates because people could take the surveys themselves on tablets and is suggestive of overall interest in the survey, while response rates in subsequent rounds are constrained by enumerator capacity. One of the June 2017 workshops was held by the Inter-American Development Bank; all other workshops were held by the World Bank. The "Total" row excludes the pilot workshops, as their data are not considered in this paper. Numbers are tentative pending final confirmation.

Table 4: Respondents by Recruitment Strategy

	Policymakers	Practitioners	Researchers
Pilot workshops	0.40	0.42	0.18
Workshops	0.36	0.33	0.31
Videoconference	0.19	0.31	0.51
Headquarters surveys	0.05	0.59	0.36
Total	0.24	0.40	0.36

This table shows the percent of respondents who could be classified as policymakers, practitioners and researchers by each recruitment strategy. Notably, the breakdown of policymakers, practitioners and researchers is not substantially different between the pilot workshops and the other workshops. The "Total" row excludes the pilot workshops, as their data are not considered in this paper.

tion, participants were asked if they understood and were only allowed to participate further if they stated that they did. Only one participant stated that they did not understand the instructions and was prohibited from continuing.

Of the 288 individuals completing the policymakers, practitioners and researchers (PPR) survey, only 41 said they had never heard of CCTs and 35 said they had heard of CCTs but did not know any studies on them. When considering how much a paper's results could improve policy decisions, we will restrict attention to those answering they either had never heard of CCTs or had heard of them but never heard of any studies on them. Of these 76 respondents, only 13 were researchers, a relatively low percentage, as we might expect. Of the 1,029 individuals completing the survey on Mechanical Turk who passed all screening tests, 530 said they had never heard of CCTs and 323 said they had heard of CCTs but did not know of any studies on them. The MTurk respondents clearly are less well-informed on average, and they may serve as a useful comparison group if one thinks that our PPR sample represents a particularly well-informed set of policymakers.

A potential limitation of using these estimates to provide a set of priors is that we may think that people who had never heard of CCTs before or who had heard of them but were not familiar with studies on them may be a selected group that might have more inaccurate priors than is typical for those considering whether or not to implement a potential new program. This is possible, though it should be noted that, in the PPR sample, those who were unaware of these studies tended to work in other areas (*e.g.* infrastructure) rather than being specifically poorly informed (*e.g.* someone who works in education in a country with a substantial CCT program who has not heard of CCTs). Further, we may expect that people with less familiarity with a program would also have wider priors and thus should update more on new information, so that the amount of updating we observe may be considered an upper bound.

3.2.1 Incentives

Policymakers, practitioners and researchers were simply offered a token gift in the workshops (chocolate or coffee costing approximately $5-$15 USD) in exchange for their time. In addition, participants were informed that at the end of the study, one response would be drawn at random and awarded an additional prize: a MacBook. We did not further incentive responses because we were concerned that policymakers in particular would fear giving a "wrong" answer, so we did not want to increase the salience of the possibility of answering "incorrectly" by offering incentives for "correct" answers. The same incentives were offered to participants at the World Bank and IDB headquarters.

For those interviews conducted over videoconference, a $15 Amazon voucher was provided, again without further conditions, along with entry to the MacBook raffle. Enumerators were trained to encourage participants who feared giving an answer by saying that there were no wrong answers and that we merely wanted to know what they thought given the information we provided - if anyone was wrong, it was our fault for how we provided information.

MTurk participants were simply offered $1.50 for the relatively long survey. We were concerned that without incentivizing thoughtful responses, participants might not put in the effort to understand and carefully answer the questions. However, we chose to implement screening questions instead as we did not want to distort responses and we thought this would provide greater comparability with the results from policymakers, practitioners and researchers. Screening questions are described in a later section.

3.2.2 Priors Data: Descriptive Statistics

This section describes the breakdown of how the respondents' priors were distributed, *i.e.* how many were normally distributed, uniformly distributed, *etc.* This section discusses this with reference to both the policymakers, practitioners and researchers sample (hereafter referred to as PPR) and the MTurk sample.

Fitting a distribution to each person's stated probability weights is complicated by the fact that we do not know *a priori* what distribution someone might have. Further, using the probability weights is time-consuming, and MTurk workers, in particular, working unsupervised, might distribute some of the probability weights with error. This is less of a concern when respondents are supervised. After a pilot with policymakers in which respondents filled out questionnaires unsupervised, we changed approaches so that an enumerator moved the slider bars based on what the respondent orally told them, with the respondent able to view what the enumerator is doing. The enumerator, in turn, would double-check that the respondent was satisfied with the distribution as they entered it before moving on to the next question.

A greater concern is that formal tests of normality such as Kolmogorov-Smirnov tests are generally inaccurate for distributions made up of a few discrete bins, as we have. While 15 bins were available, most respondents' estimates fell into 7 or fewer bins. Given these issues, we simply ask: what is the overall shape of the distributions that people reported? To do this in a transparent way and account for error in moving the slider bars in the survey when respondents are unsupervised, we start by considering how many times the probability weights appeared to increase or decrease as we move from one bin to the next sequentially and using this to classify distributions into different types. For example, the bins that we use for most of the study range from -5 to -4 percentage points, -4 to -3 percentage points, *etc.*, all the way to 9 to 10 percentage points. We might observe someone putting some weight in the 1-2 bin, more weight in the 2-3 bin, and less weight in the 3-4 bin, with zero weight on all other bins. This would look like a pattern of: increase (from zero weight to some weight) - increase - decrease - decrease (from some weight to zero weight). Alternatively, we might observe someone putting a lot of weight in the 1-2 bin, some weight in the 2-3 bin, and a lot of weight in the 3-4 bin, a pattern of: increase - decrease - increase - decrease. The first distribution could be normal (though it might not be) but the second is definitely not normal, and could instead perhaps be bimodal. In the first case, the weights start at zero,

Table 5: Simple Classification Scheme for Distributions

Bimodal	increase - decrease - increase - decrease
Decreasing	decrease
Increasing	increase
Normal	increase - decrease
Normal, left tail	decrease - increase - decrease
Normal, right tail	increase - decrease - increase
Uniform	increase - decrease, with equal weights

This table describes a simple classification scheme for observed patterns of probability weights. The scheme is meant to give a broad, descriptive overview of the different kinds of distributions that are reported. In particular, it can help to quickly summarize which of the distributions appear to not be normal. "Normal, left tail" denotes a distribution that would otherwise be classified as normal but starts with some extra probability weight in the left tail such that we observe a decrease in weights before the weights begin to increase again; "Normal, right tail" has an analogous meaning.

hit a single peak, and decrease to zero. In the second, they weights start at zero, hit two peaks, and end at zero. Table 5 lists several common distributions (and some less common ones) and what they might require of the overall shape of the distribution, according to this basic classification scheme. To simplify, we note only when there is a change in direction from "increase" to "decrease" or vice versa. For example, the case previously described as increase - increase - decrease - decrease can be more simply described as increase - decrease.

A special case here is the uniform distribution. If we observe that someone places equal weights in multiple bins, we consider that a uniform distribution rather than a normal distribution. Conservatively, we consider this to be uniform even in the (very rare) case in which someone puts all the weight in one bin, when the distribution could be normal or any number of other distributions. This pathological case will have no bearing on the main results, which focus on those cases in which respondents reported normally distributed priors.

Again, we are aware that even if a set of priors followed a pattern with one peak and was not a uniform distribution, it might not be a normal distribution. Still, this classification scheme can help to quickly describe many of the kinds of distributions that we might observe and qualitatively summarize them.

As weights may be slightly noisy, we only count an increase if weights increase by 5 or

more from one bin to the next, and we only count a decrease if weights decrease by 5 or more from one bin to the next. For example, the bins that we use for most of the study range from -5 to -4, -4 to -3, *etc.*, all the way to 9 to 10. We might observe someone putting a weight of 3 on the 1-2 bin; a weight of 9 on the 2-3 bin; a weight of 20 on the 3-4 bin; a weight of 33 on the 4-5 bin; a weight of 20 on the 5-6 bin; a weight of 7 on the 6-7 bin; and a weight of 8 on the 7-8 bin, with weights of 0 on all other bins. If we look at this distribution, it looks plausibly normal, but that last bin has more weight than we might expect. Perhaps the respondent was tired, or perhaps they wanted to make the weights add to 100 and put the excess weight in that last bin.[9] Again, we would drop that respondent's distribution to be conservative.

Table 6 provides the distribution of priors for policymakers, researchers and practitioners, as categorized in this basic descriptive scheme. 29% reported priors that took a clearly non-normal distribution when observing the pattern of increases and decreases in the weights across bins. We expect that some appear to have had "increasing" priors simply because they wanted to guess higher values than were available in the slider tool (recalling that the highest bin allowed for an increase in enrollment rates by 10 percentage points).

Of 1,675 MTurk respondents[10], 1029 passed the "screening" questions. One set of screening questions, described in detail elsewhere (Vivalt and Coville, 2017), required respondents to not update very differently across several questions; 1183 passed this requirement. There was also one screening question which provided them with a set of pre-filled slider bars and asked them to adjust the slider bars to reflect their beliefs given new information; as participants were not required to adjust the bars at all, but the question was designed such that reasonable people would adjust their estimates upwards, we dropped anyone who did not adjust any of the bars, considering them potentially too inattentive or uninterested in

[9] All responses were normalized such that weights added to 100, and respondents were informed that their answers would be normalized, but some insisted on making the weights add to 100 regardless.

[10] We accidentally gathered slightly more data than the 1,600 responses initially planned, as a few more people answered the survey than filled in a survey code on MTurk within the alloted time, such that the HIT was not counted and was re-offered to other participants.

Table 6: Classification of Reported Priors

	PPR		MTurk	
Rough Category	Total	Percent	Total	Percent
Bimodal	0	0.00	14	0.02
Decreasing	0	0.00	12	0.01
Increasing	7	0.13	72	0.08
Normal	37	0.71	706	0.83
Normal, left tail	0	0.00	1	0.00
Normal, right tail	0	0.00	7	0.01
Uniform	8	0.15	33	0.04
Other	0	0.00	8	0.01
Total	52	1.00	853	1.00

This table shows the basic shape of the priors for each of the PPR and MTurk samples, restricting attention to those who stated they had either never heard of CCTs or had heard of them but never heard of any studies on them.

putting effort into their responses. Restricting attention to those who claimed to have not heard of conditional cash transfer programs or of any studies on conditional cash transfer programs, we are left with a set of 853 respondents.

Of these, 17% reported priors that took a clearly non-normal distribution when observing the pattern of increases and decreases in the weights across bins. Table 6 illustrates. Of those responses that were plausibly normally distributed, we fit their distribution to the closest normal distribution by simulation.

4 Method

Policymakers are assumed to Bayesian update as each new study arrives, using the Bayesian hierarchical model. We consider three sets of potential priors they might have: completely uninformative (uniform) priors; slightly informed, normally distributed priors based on the mean and standard error of a study drawn at random from within each intervention-outcome combination; and normally distributed priors drawn at random from a set of real priors.

The advantages of using uninformative priors are, firstly, that this most closely follows the literature and, secondly, that this case leads to the greatest benefits upon receipt of a new study. To make a perhaps more realistic set of normally distributed priors, we draw one result from the set of all results within an intervention-outcome combination and use that result to update the uninformed prior, resulting in normally distributed, slightly informed priors.[11] The advantage of this second set of priors is that they represent a partially-informed set of priors, and we may think that completely uninformed priors are unrealistic; the drawback of this approach is that these priors may give policymakers too much credit and be more accurate than their priors would typically be, so as to underestimate the benefits of impact evaluation. To address this concern, we leverage a third set of priors directly elicited from poorly-informed policymakers, practitioners and researchers or, alternatively, MTurk workers. These priors were provided for the case of CCTs, and we restrict attention to the priors of those who stated they had never heard of the results of any study on that type of program; we may imagine that those who are better-informed and consequently have narrower priors would update less on the new information, so these estimates provide an upper bound. It remains possible that respondents had heard something about the program's effects without hearing of any study, but further restricting attention would limit us to a very small sample of priors.

The results data that the policymakers in our model will use to update is randomly drawn from the set of real results data, within each intervention-outcome combination. We will repeatedly sample from these data to form a variety of permutations of the order in which these studies could arrive.

Using the priors and the new data, we will follow the estimation strategy described in section 2.3 to estimate $\widehat{\mu}$ (or $\widehat{\theta}_i$ for the mixed model). We will then calculate B for each draw. It is important to emphasize that all analyses, from the estimation of parameters to the estimation of B, are conducted within intervention-outcome combinations (*e.g.* the

[11] We subsequently exclude that study from the set of possible "real" new results.

effect of CCT programs on enrollment rates).

Note that in order to calculate the benefits, B, to an impact evaluation, we need to know the actual best choice; in other words, we need to know the mean true effect μ in the random-effects model and the true effect θ_i in the mixed model as compared to θ^*. We will assume that the program's μ can be approximated by our estimate for μ using all data, $\widehat{\mu}_N$, for large enough N and that, similarly, estimates of θ_i using all data, $\widehat{\theta}_{iN}$, will converge to the true value of θ_i for each θ_i. We will then assume that the outside option θ^* would have 50-90% of the effect of the program under consideration, in alternative specifications.

The results section will first present average estimates of $B_{n,n+1}$ for moving from the n^{th} to the $n+1^{th}$ study within intervention-outcome in a random-effects model using uninformed priors. Following this, we will present results from simulations that show how $B_{n,n+1}$ varies by σ_i^2 and τ^2, for various θ^* and μ. Since $B_{n,n+1}$ is fully specified by these parameters, this can give some insight as to what we might observe with different data. We then turn to discussing how results would change in a mixed model or using different priors. Finally, we separately consider the case in which a policymaker can make a decision on their own, drawing a prior randomly from the set of real priors, and the case in which policymakers make decisions in small groups of size n=2,...,10, following a majority voting rule.[12]

5 Results Using Simulated Priors

5.1 Uninformative Priors

The first case we will consider is the case of uninformative priors. We will estimate marginal benefits to an additional study and then simulate marginal benefits under different parameter values and under a mixed model.

[12]Ties are broken randomly.

5.1.1 Estimated Marginal Benefits

The marginal benefits to an additional study under uninformative priors are empirically calculated by considering the order in which the studies within an intervention-outcome combination could have arrived and whether a study would have changed the policymaker's decision in a positive way or in a negative way if they had arrived in that order. As discussed, this requires an assumption about θ^* relative to μ. We assume $\theta^* = 0.5\mu$ and $\theta^* = 0.9\mu$ under alternative specifications. On the one hand, one might be interested in the benefits when $\theta^* = 0.5\mu$, as this yields the difference between θ^* and μ we might expect if results were not correlated within outcomes across different interventions: the mean effect size in the results data is 0.12,[13] and if one were to simply randomly sample two effect sizes from the distribution, the median absolute difference between them would be 0.08. We could imagine, for example, that a policymaker might face a true θ_i of 0.16 and an outside option of $\theta^* = 0.08$. On the other hand, $\theta^* = 0.9\mu$ may be reasonable as interventions targeting the same outcome frequently obtain similar effects and if results were far apart it is more likely a policymaker would know which was the better option without a study.

Results for $\theta^* = 0.5\mu$ are summarized in Table 7.[14] The median marginal benefit across intervention-outcome combinations from moving from 1 to 2 studies is 0.0012, or about 1%; for the top 20%, it is more than 0.0076 (6.5%). By the 10th study, much of the benefits have been realized. The top 20% of intervention-outcome combinations show a benefit of only 0.0019 (1.6%) when moving from the 10th to the 11th study. Further, this is if anything upwards-biased, as we might imagine that one is more likely to do that many studies if one expects to find divergent results.

Benefits are lower when we assume $\theta^* = 0.9\mu$.[15] Here, the bottom 20% of intervention-outcome combinations show marginal benefits of less than 0.0001 (0.1% for the 2nd study);

[13] Throughout, "effect size" refers to the standardized effect sizes, as is conventional in the meta-analysis literature, *e.g.* 0.12 represents a change in the outcome variable of 0.12 standard deviations.

[14] Table 9 in the Appendix provides full results for each intervention-outcome combination.

[15] Table 10 in the Appendix provides full results for each intervention-outcome combination.

Table 7: Marginal Benefits of an Additional Study

	$B_{1,2}$	$B_{5,6}$	$B_{10,11}$
$\theta^* = 0.5\mu$			
20th percentile	0.0001	0.0001	0.0000
40th percentile	0.0006	0.0006	0.0002
60th percentile	0.0018	0.0010	0.0004
80th percentile	0.0076	0.0024	0.0019
$\theta^* = 0.9\mu$			
20th percentile	0.0000	0.0000	0.0000
40th percentile	0.0001	0.0001	0.0001
60th percentile	0.0003	0.0002	0.0003
80th percentile	0.0008	0.0007	0.0005

This table shows the calculated marginal benefits, $B_{n,n+1}$, of moving from study n to study $n+1$, assuming that $\theta^* = 0.5\mu$ or $\theta^* = 0.9\mu$. To generate this figure, we form each possible order of studies' results within each intervention-outcome, calculate whether, in each case, the $n+1^{th}$ study would be pivotal and in which direction, and then take the expected value of the benefits across all the different possible ways to move from n to $n+1$ studies within that intervention-outcome. All benefits are in terms of effect sizes and all $B_{n,n+1}$ calculations are done on whichever intervention-outcomes have at least $n+1$ studies, meaning that the columns are not strictly comparable to each other as different intervention-outcome combinations could be included in each. As a point of reference, the typical effect size of a study is 0.12.

for the top 20%, it is more than 0.0008 (0.7%). The median benefit is 0.0002, or approximately 0.2% of the typical effect size in the data. Again, the more studies that are done, the lower the benefits are.

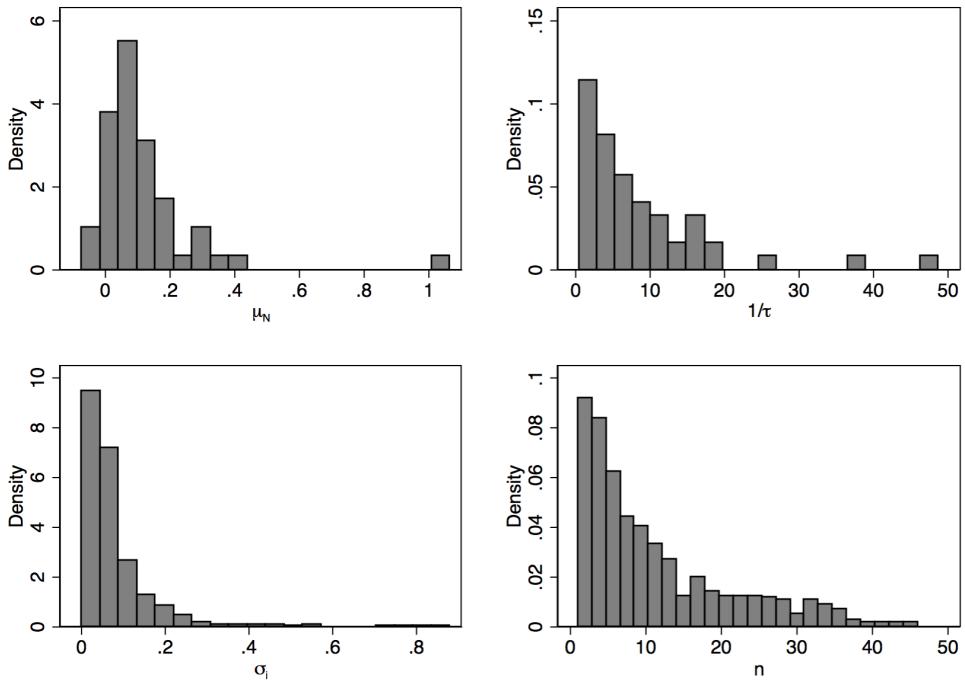
Table 11 in the Appendix presents results restricting attention to those intervention-outcome combinations with N $\geqslant$ 10, for which $\widehat{\mu}_N$ may better approximate μ. Results are broadly comparable.

5.1.2 Simulated Marginal Benefits

To further investigate the issue, we simulate B for different parameter values. The relevant factors are, as discussed, μ, θ^*, τ^2, σ_i^2 and n.

Again, the choice of μ and θ^* is clearly important. As before, we will assume $\theta^* = 0.5\mu$ or $\theta^* = 0.9\mu$. We also need reasonable values of τ^2, σ_i^2 and n. Figure 1 shows the density of $1/\tau$, σ_i and n in the data. Note that $1/\tau$ is calculated on the full data within each

Figure 1: Distribution of Parameters



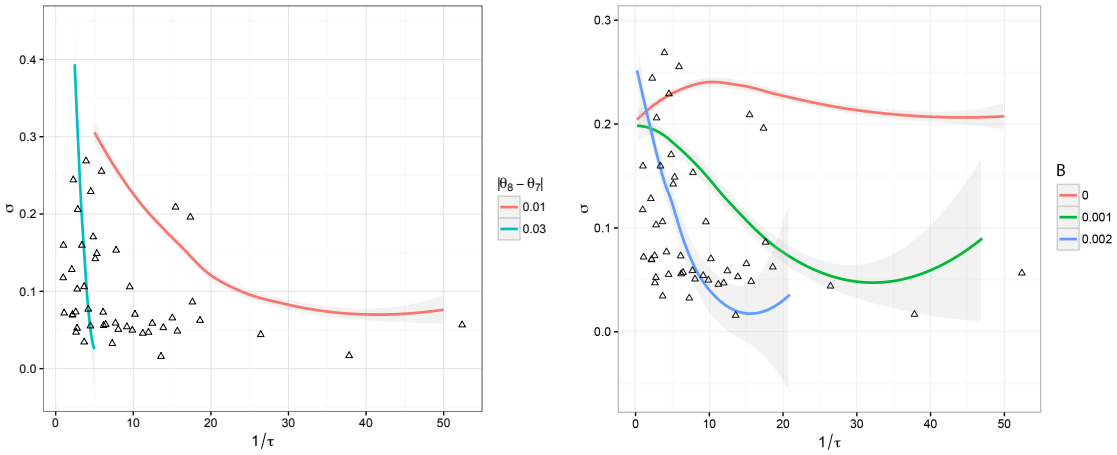
This table shows the distribution of parameter values estimated from hierarchical Bayesian meta-analyses within each intervention-outcome combination.

intervention-outcome combination. σ_i is supplied by each study. For n, the first study on a topic is counted as $n = 1$, the second, $n = 2$, *etc.*

The benefit of using simulations is that we can explore what would happen with different parameter values. Note that while we have σ_i from each study, to consider the benefits to a particular group's aggregate σ_i we need a way to aggregate the σ_i within a group. We will use the average σ_i within a group, denoted σ. To explore what would happen with different parameter values, we first generate a set of σ and τ for reasonable values of each - drawing 50 equispaced values between 0.01 to 0.4 for σ and a sequence increasing in 0.15 increments between 0.2 to 50 for $1/\tau$. For each σ and τ, we simulate (500 times) a data set with $n + 1$ data points, generating θ_i, Y_i, $\widehat{\mu}_n$, and $\widehat{\mu}_{n+1}$. This enables us to know when the $n + 1^{th}$ study is pivotal and what the benefits are when it is. Figure 2 shows how much estimates of $\widehat{\mu}_n$ change and the marginal benefits to doing an additional study for different values of σ and τ. Figure 2 assumes that $\theta^* = 0.5\mu$ and 7 impact evaluations have been done to date (pooling across all studies, the number of studies previously done on an intervention-outcome combination is 7). The triangles overlaid on top of the simulation results represent the actual $1/\tau$ and average σ_i we observe within each intervention-outcome combination. Most are clustered around marginal benefits of approximately 0.001-0.002, an improvement of 1-2%. It does not appear that changes in τ and σ would substantially affect B.

Additional assumptions could further improve the situation. For example, perhaps the policymaker is not completely uninformed as to whether an additional impact evaluation would be pivotal and elects to only do an impact evaluation when the chance that it would be pivotal is high. Spillover effects to other policymakers could also be considered, as a study's results are a public good, multiplying the benefits while holding the costs constant. The benefits here are a simple function of how many others could benefit: if we consider that the median intervention-outcome combination has 7 studies, this suggests that at least 6 other projects might benefit from the first study's results, so all estimates should be scaled accordingly, with the first few studies again having a proportionately larger impact than

Figure 2: Marginal Benefits to Doing An Additional Study



The figure on the left shows the hypothetical improvements in the estimation of μ when doing an additional study, for various $1/\tau$ and average σ_i. Each contour represents a particular amount by which the estimation is improved. Triangles represent the $1/\tau$ and average σ_i in the data, while curves are drawn using simulated data. We assume the additional study would be the 8th study, as the median number of studies on a particular topic is 7 in the data. The figure on the right shows the hypothetical benefits the policymaker would realize in terms of policy outcomes, given that the improvement in the estimation of μ is only pivotal some of the time (and some of the time is pivotal but misleading, by chance). For these simulations, the outside option is assumed to have the value $\theta^* = 0.5\mu$, and the true value of μ is assumed to be 0.12 (the mean effect size).

later studies.

In summary, there are likely substantial gains to be realized in shifting our attention to those areas with fewer studies and greater initial uncertainty. We cannot easily advise researchers to find those opportunities with a particular θ^* and μ, since if these were known there would be no point to the study; however, it does make sense to prioritize those cases in which μ could be revealed to be much different from θ^*, *i.e.* when there is a lot of uncertainty. Notably, we find that the greatest policy benefits accrue to a very small number of studies. This is similar to the power law observed in cost-effectiveness analyses (Jamison *et al.*, 2006).

5.1.3 Simulated Marginal Benefits Under a Mixed Model

In the case that we can explain much of the observed heterogeneity, the marginal benefit of an impact evaluation could be higher. Suppose that instead of Equation (4), we believed:

$$Y_i = \theta + \beta X_i + \eta_i + \varepsilon_i \tag{14}$$

where X_i represents an explanatory variable.

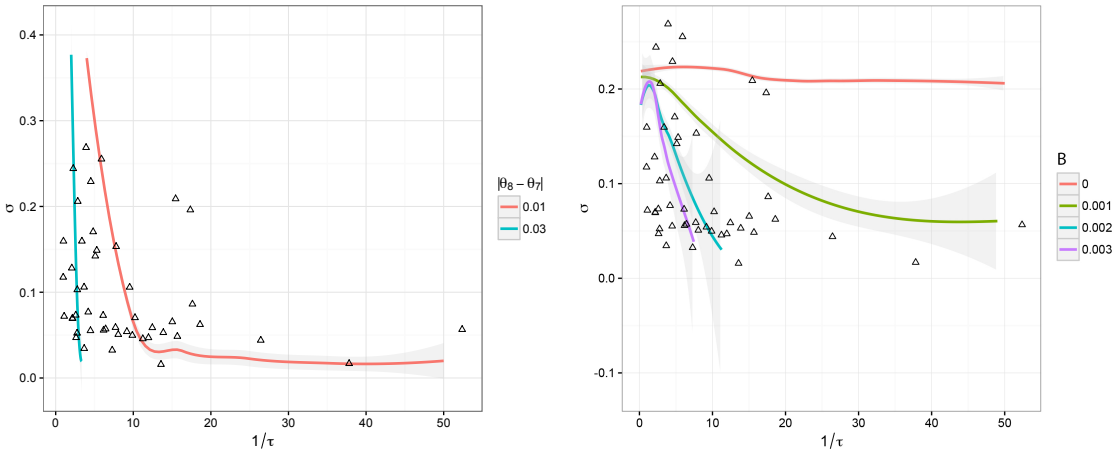
Then each impact evaluation would give us information not just about μ but also about β. A policymaker's estimate of θ_i would be $\mu + \beta X_i$ for the X_i in their context. The equation for the marginal benefits would now be:

$$B = p^+ \cdot (|\theta_i - \theta^*|) - p^- \cdot (|\theta_i - \theta^*|) \tag{15}$$

where $p^+ = \mathrm{P}(\widehat{\mu}_{n+1} + \widehat{\beta}_{n+1} X_i > \theta^*, \widehat{\mu}_n + \widehat{\beta}_n X_i < \theta^*)$ and $p^- = \mathrm{P}(\widehat{\mu}_{n+1} + \widehat{\beta}_{n+1} X_i < \theta^*, \widehat{\mu}_n + \widehat{\beta}_n X_i > \theta^*)$ for the case that $\mu + \beta X_i > \theta^*$.

In the data, we do not observe a variable that explains much of the heterogeneity in treatment effects for all intervention-outcome combinations (Vivalt, 2017a). Instead, variation seems to be better-explained using variables unique to each intervention-outcome combination. Vivalt (2017a) considers the effect of CCTs on enrollment rates and finds explanatory

Figure 3: Marginal Benefits to Doing An Additional Study Under a Mixed Model



The figure on the left shows the hypothetical improvements in the estimation of μ when doing an additional study, for various $1/\tau$ and average σ_i. Each contour represents a particular amount by which the estimation is improved. Triangles represent the $1/\tau$ and average σ_i in the data, while curves are drawn using simulated data. We assume the additional study would be the 8th study, as the median number of studies on a particular topic is 7 in the data. The figure on the right shows the hypothetical benefits the policymaker would realize in terms of policy outcomes, given that the improvement in the estimation of μ is only pivotal some of the time (and some of the time is pivotal but misleading, by chance). For these simulations, the outside option is assumed to have the value $\theta^* = 0.5\mu$, and the true value of μ is assumed to be 0.12 (the mean effect size).

variables yielding an R^2 of approximately 0.5 in the best case scenarios. We will optimistically assume that an explanatory variable could be found for each intervention-outcome such that an OLS regression of Y_i on X_i within each intervention-outcome would yield an R^2 of 0.5.

To simulate this, we begin with the same Y_i used in the simulations of the random-effects model and generate X_i under this constraint by adding noise. τ^2 and σ_i^2 are then re-estimated, as they now take on different values.[16] Figure 4 shows the results for the case that $X_i = \bar{X}$.

As we can see, the marginal benefits of a new study for a given σ_i and τ are higher but not by much if we can model the heterogeneity in treatment effects.

[16]Otherwise, with Y_i, τ^2 and σ_i^2 specified, $\beta X_i = Y_i - N(0, \tau) - N(0, \sigma_i)$ and we cannot control R^2.

5.2 Normally Distributed Priors

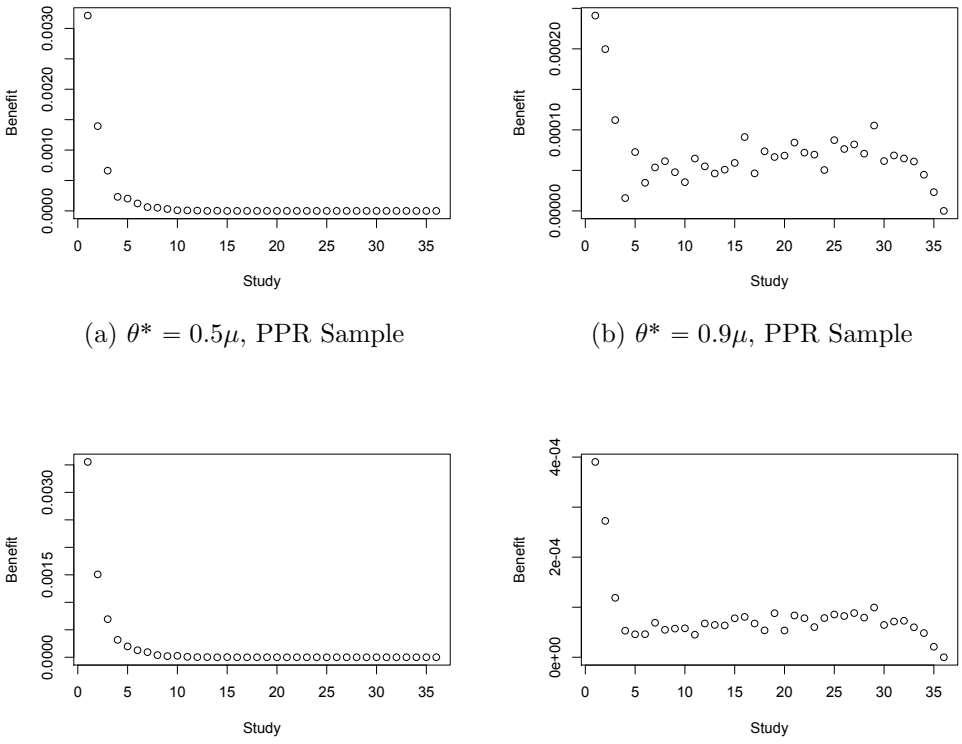
What if, rather than being completely uninformed, policymakers were slightly informed, with normally distributed priors? For example, building upon the case of policymakers having uninformative priors, we might perhaps take the uninformative, uniform priors and add a randomly-selected study via Bayesian updating to form the normally distributed priors. This results in estimates exactly comparable to the previous case, except that the previous benefit $B_{i,j}$ is now $B_{i-1,j-1}$. In other words, all benefits are slightly reduced.

Perhaps this case is too charitable to policymakers. If they had less well-informed priors, B would be higher. We now turn to exploiting a set of real priors. These priors add realism but restrict attention to the case of the effects of CCTs on enrollment rates.

6 Results Using Real Priors

6.1 Single Dictator

As in the previous section, a single policymaker decides which program to implement. Their prior is randomly drawn from the set of real priors and updated each period with new information. Figure 4 illustrates the average marginal benefits to an additional study by the study number, using the PPR or MTurk samples, respectively. This figure suggests not much changes from the earlier estimates: the marginal impacts of an additional study in any one given context are still quite small, ranging from approximately 0.1-0.3 percentage points for the first two studies in the case that $\theta^* = 0.5\mu$ and 0.02-0.04 percentage points for the first two studies in the case that $\theta^* = 0.9\mu$, down to less than 0.01 percentage points in either case. The median CCT in this data set improves enrollment rates by 5 percentage points, so these numbers reflect an improvement in outcomes by 0.4-6%, in line with the earlier estimates based on uninformative priors. Of course, if one imagines the early studies may have informed several dozen programs, this could increase the benefits substantially.



(a) $\theta^* = 0.5\mu$, PPR Sample

(b) $\theta^* = 0.9\mu$, PPR Sample

(c) $\theta^* = 0.5\mu$, MTurk Sample

(d) $\theta^* = 0.9\mu$, MTurk Sample

Figure 4: Marginal Benefits of an Additional Study Under Single Dictator, Using Real Priors

This table shows the average calculated marginal benefits, $B_{i,j}$, of an additional study, assuming that $\theta^* = 0.5\mu$ or $\theta^* = 0.9\mu$ and that there is a single dictator making the decision, and using real priors and results data. To generate this figure, we randomly draw a prior from the set of real priors data and use this as the prior of the dictator making a decision. The individual is then assumed to Bayesian update based on the new information from a study. The study's results are also randomly drawn from a set of real results data. μ, and hence θ^*, is estimated as the mean hierarchical Bayesian meta-analysis point estimate, using data from all studies within the intervention-outcome combination and 10,000 simulations.

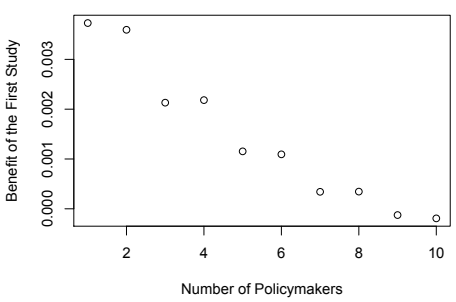
6.2 Majority Voting Rule

The case in which multiple policymakers jointly make a decision may be of separate interest. Here, we will consider the case in which a group of 2-10 policymakers jointly determines which program to implement by majority vote. As in the previous scenario, each policymaker's prior is randomly drawn from the set of real priors and updated each period with new information from a randomly-drawn study. A vote is taken at the beginning, before any studies have occurred, and after every study thereafter. Figure 5 illustrates the average marginal benefits to the first study under this scenario, by the number of decision makers. The marginal benefit of the first study actually declines with the number of decision makers for most cases, although for the case of $\theta^* = 0.9\mu$ and the relatively poorly-informed MTurk sample, the benefits increase with the number of decision makers. This makes intuitive sense: whether the benefits of an additional study rise or fall with the number of decision makers depends on the decision makers' priors and the true effects of the programs.

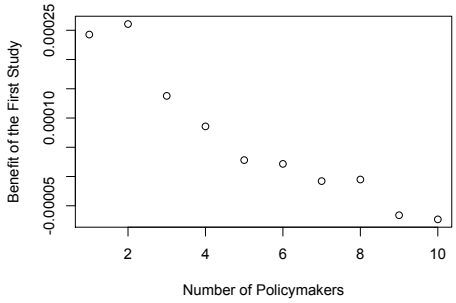
It may be instructive to consider how much policy decisions might be improved by having multiple policymakers vote in the absence of any evidence. In this case, policymakers would simply be leveraging the "wisdom of the crowds" and could be making a better judgment together than they would likely make individually. Figure 6 illustrates the benefits of an additional decision maker when no information from a study is available. Notably, for the case in which $\theta^* = 0.5\mu$, adding a decision maker has decreasing marginal returns. For the $\theta^* = 0.9\mu$ case, additional decision makers have ambiguous impact.[17]

Careful examination of Figures 4-6 shows that increasing the number of decision makers can sometimes lead to greater benefits in policy outcomes than running an additional study.

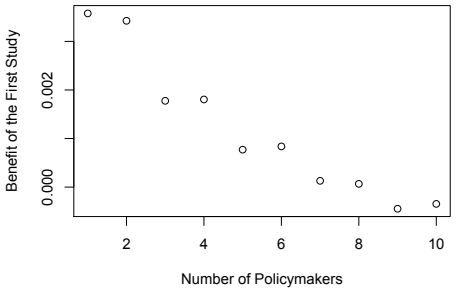
[17]The argument as to why we should expect that even-numbered decision makers have no impact, as we see in this Figure, can be sketched out as follows. Suppose that we are moving from one decision maker to two. Then there is some probability that the second one will disagree with the first one, and which program is selected will be randomly determined. If the second decision maker does not disagree with the first decision maker, the addition of the second decision maker has no impact on the decision. If they do disagree with the first decision maker, it is equally likely that the first decision maker was wrong and the second decision maker was right as it is that the first decision maker was right and the second decision maker was wrong. Therefore, in expectation, they have no impact one way or another on the quality of the decision, and the argument can be extended to all cases of moving from an odd number of decision makers.



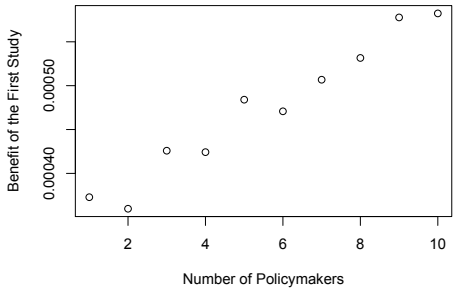
(a) $\theta^* = 0.5\mu$, PPR Sample



(b) $\theta^* = 0.9\mu$, PPR Sample



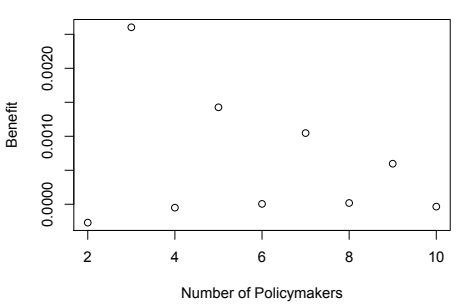
(c) $\theta^* = 0.5\mu$, MTurk Sample



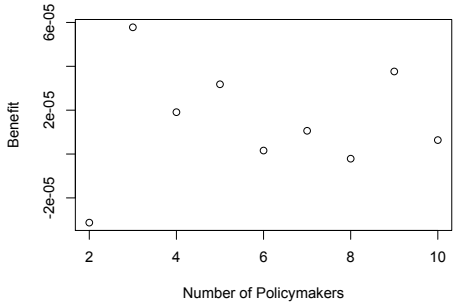
(d) $\theta^* = 0.9\mu$, MTurk Sample

Figure 5: Marginal Benefits of the First Study Under Majority Voting Rule, by Number of Decision Makers

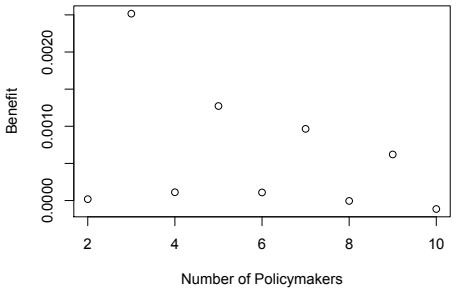
This table shows the average calculated marginal benefits, $B_{0,1}$, of the first study on a topic, assuming that $\theta^* = 0.5\mu$ or $\theta^* = 0.9\mu$ and using a majority voting rule in conjunction with real priors and results data. To generate this figure, we randomly draw a set of N priors from the set of real priors data and use these as the priors of N policymakers making a decision under a majority voting rule. These individuals are then assumed to Bayesian update based on the new information from a study. The study's results are also randomly drawn from a set of real results data. μ, and hence θ^*, is estimated as the mean hierarchical Bayesian meta-analysis point estimate, using data from all studies within the intervention-outcome combination and 10,000 simulations.



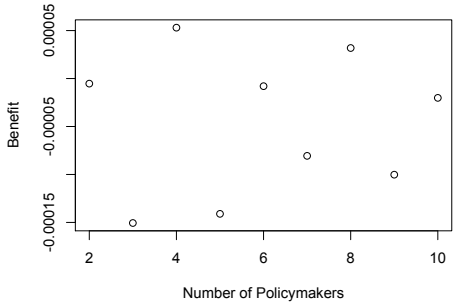
(a) $\theta^* = 0.5\mu$, PPR Sample



(b) $\theta^* = 0.9\mu$, PPR Sample



(c) $\theta^* = 0.5\mu$, MTurk Sample



(d) $\theta^* = 0.9\mu$, MTurk Sample

Figure 6: Marginal Benefits of an Additional Decision Maker Under Majority Voting Rule and No Study Information

This table shows the average calculated marginal benefits of adding a decision maker when no study information is available, assuming that $\theta^* = 0.5\mu$ or $\theta^* = 0.9\mu$ and using a majority voting rule in conjunction with real priors and results data. To generate this figure, we randomly draw a set of N priors from the set of real priors data and use these as the priors of N policymakers making a decision under a majority voting rule. These individuals then vote on which program to support, with ties broken randomly. μ, and hence θ^*, is estimated as the mean hierarchical Bayesian meta-analysis point estimate, using data from all studies within the intervention-outcome combination and 10,000 simulations.

For example, moving from N=1 to N=10 decision makers in Figure 5d provides the majority of the benefits of conducting a second study holding the number of decision makers constant at N=1 in Figure 4d. Further, if 10 decision makers voted on which program to pursue in the absence of any study information for the case in which $\theta^* = 0.5\mu$ (Figure 6a or Figure 6c), they would obtain better outcomes than a single dictator would obtain if given information from a first study (Figure 4a or 4c).

7 Discussion

The model in this paper is intentionally simple. It assumes policymakers care about evidence and that they update correctly based on that evidence.

There is another respect in which the results are somewhat optimistic. One may criticize the lumping together of similar but not exactly identical interventions. Context also varies across studies, and policymakers may prefer to restrict attention to a subset of studies that are closer to their setting. If we did restrict attention to more similar interventions or contexts, the point estimates of the studies included might be better predictors of the target θ_i and these more similar studies might have a lower inter-study variance τ^2. However, if this is true, then the policymaker would have less incentive to fund another study in their context; as τ^2 falls, so does the value of another study.

In fact, the model implies a trade-off between learning about what works best in a given context and external validity. In the model, a study done in a particular setting is perfectly informative of what will happen in that setting if the program were to be replicated, apart from sampling variance. One learns more from a study when τ^2 is high, however, if τ^2 is high, then that study will itself be less informative to other policymakers in different contexts in the future.

Finally, the results highlight the extent to which better decisions could be made simply by democratizing the decisionmaking process: decisions made by majority vote by a larger

number of policymakers can be better than decisions made by a single policymaker, due to wisdom of the crowds, and adding decision makers can result in better policy outcomes than conducting a study. However, this will depend on parameter values; priors matter.

8 Conclusion

How much impact evaluation results can inform policy in other settings is an important topic. We model policymakers as having a prior about the effects of a program and updating that when new information comes out. This enables us to estimate the marginal benefit of an additional impact evaluation. The improvements in policy decisions that would be realized depend on the value of the outside option. Picking alternative values of $\theta^* = 0.5\mu$ or $\theta^* = 0.9\mu$, back-of-the-envelope calculations suggest typical improvements of about 1%-6.5%, decreasing as more impact evaluations are completed.

Since explanatory models could help provide policymakers with more guidance, we discussed how estimates might change under a mixed model. The marginal benefits of impact evaluation were improved, but not substantially. The main results are largely driven by the relatively small differences observed between interventions, so explaining more of the heterogeneity in treatment effects is unlikely to have a substantial effect.

Finally, we leveraged real priors data, considering the case of the effect of CCTs on enrollment rates. Here, the information provided by the first study led to an estimated 0.02-0.3 percentage point (0.4-6%) improvement in enrollment rates in a given context, depending on the value of the outside option. If multiple policymakers jointly arrived at a decision by majority vote, this estimated benefit increased to 0.03-0.4 percentage points (0.6-8%).

Despite these relatively small estimates, a study may still be worthwhile if multiple policymakers could take advantage of the information it would provide, as impact evaluations are public goods. Further, a small change in effects might still be important if large numbers of people are targeted by the program. This paper focuses on the impact in a given setting but

these estimates can be scaled up depending on how often they are used in practice. However, as we saw, the more useful a paper is in one particular context due to heterogeneity in true underlying treatment effects, the less useful that paper will be in other contexts.

Overall, the results suggest that greater attention be paid to characteristics of studies that help determine whether an additional study would be worthwhile in order to maximize their benefits. Novel programs that have the potential to show particularly large effects are the most promising candidates for evaluation, and carefully explaining heterogeneity in treatment effects for a particular intervention would also help improve evidence-based policymaking. Finally, the results highlight that forming decisions as a group can be more important for making a good policy decision than running an additional study, suggesting that an increased focus on decision making processes could yield substantial benefits.

References

AidGrade (2013). "AidGrade Process Description", http://www.aidgrade.org/methodology/.processmap-and-methodology, March 9, 2013.

AidGrade (2015). "AidGrade Impact Evaluation Data, Version 1.2".

Bandiera, Oriana, Greg Fischer, Andrea Prat and Erina Ytsma (2015). "Building Evidence from Multiple Studies: The Response to Incentive Pay", working paper.

Borenstein, Michael *et al.* (2009). Introduction to Meta-Analysis. Wiley Publishers.

Brodeur, Abel *et al.* (2012). "Star Wars: The Empirics Strike Back", working paper.

Cartwright, Nancy (2007). Hunting Causes and Using Them: Approaches in Philosophy and Economics. Cambridge: Cambridge University Press.

Cartwright, Nancy (2010). "What Are Randomized Controlled Trials Good For?", Philosophical Studies, vol. 147 (1): 59-70.

Casey, Katherine, Rachel Glennerster, and Edward Miguel (2012). "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan." Quarterly Journal of Economics, vol. 127 (4): 1755-1812.

Cohen, Jacob (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Coville, Aidan and Eva Vivalt (2016). "How Often Should We Believe Positive Results?", working paper.

Deaton, Angus (2010). "Instruments, Randomization, and Learning about Development." Journal of Economic Literature, vol. 48 (2): 424-55.

Dehejia, Rajeev, Cristian Pop-Eleches and Cyrus Samii (2015). "From Local to Global: External Validity in a Fertility Natural Experiment", working paper.

Duflo, Esther, Pascaline Dupas and Michael Kremer (2012). "School Governance, Teacher Incentives and Pupil-Teacher Ratios: Experimental Evidence from Kenyan Primary Schools", NBER Working Paper.

Duflo, Esther, Rachel Glennerster and Michael Kremer (2008). "Using randomization

in development economics research: a toolkit", in P. Schultz and J. Strauss, eds., Handbook of Development Economics. Amsterdam: North Holland.

Evans, David and Anna Popova (2014). "Cost-effectiveness Measurement in Development: Accounting for Local Costs and Noisy Impacts", World Bank Policy Research Working Paper, No. 7027.

Gechter, Michael (2015). "Generalizing the Results from Social Experiments: Theory and Evidence from Mexico and India", working paper.

Gelman, Andrew *et al.* (2013). Bayesian Data Analysis, Third Edition, Chapman and Hall/CRC.

Hedges, Larry and Therese Pigott (2004). "The Power of Statistical Tests for Moderators in Meta-Analysis", Psychological Methods, vol. 9 (4).

Higgins, Julian PT and Sally Green, *(eds.)* (2011). Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0 [updated March 2011]. The Cochrane Collaboration. Available from www.cochrane-handbook.org.

Higgins, Julian PT *et al.* (2003). "Measuring inconsistency in meta-analyses", BMJ 327: 557-60.

Higgins, Julian PT and Simon Thompson (2002). "Quantifying heterogeneity in a meta-analysis", Statistics in Medicine, vol. 21: 1539-1558.

Independent Evaluation Group (2012). "World Bank Group Impact Evaluations: Relevance and Effectiveness", World Bank Group.

Innovations for Poverty Action (2015). "IPA Launches the Goldilocks Project: Helping Organizations Build Right-Fit M&E Systems", http://www.poverty-action.org/goldilocks.

Jadad, A.R. *et al.* (1996). "Assessing the quality of reports of randomized clinical trials: Is blinding necessary?" Controlled Clinical Trials, 17 (1): 112.

Jamison, Dean T., Joel G. Breman, Anthony R. Measham, George Alleyne, Mariam Claeson, David B. Evans, Prabhat Jha, Ann Mills, Philip Musgrove, *eds.* (2006). "Disease

Control Priorities in Developing Countries, Second Edition". Washington, DC: World Bank and Oxford University Press.

Meager, Rachel (2015). "Understanding the Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of 7 Randomised Experiments", working paper.

Millennium Challenge Corporation (2009). "Key Elements of Evaluation at MCC", presentation June 9, 2009.

Rodrik, Dani (2009). "The New Development Economics: We Shall Experiment, but How Shall We Learn?", in What Works in Development? Thinking Big, and Thinking Small, ed. Jessica Cohen and William Easterly, 24-47. Washington, D.C.: Brookings Institution Press.

Rubin, Donald (1981). "Estimation in Parallel Randomized Experiments", Journal of Educational and Behavioral Statistics, vol. 6(4).

Saavedra, Juan and Sandra Garcia (2013). "Educational Impacts and Cost-Effectiveness of Conditional Cash Transfer Programs in Developing Countries: A Meta-Analysis", CESR Working Paper.

Shadish, William, Thomas Cook and Donald Campbell (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton Mifflin.

Tierney, Michael J. *et al.* (2011). "More Dollars than Sense: Refining Our Knowledge of Development Finance Using AidData", World Development, vol. 39.

Vivalt, Eva and Aidan Coville (2017). "How Do Policymakers Update?", working paper.

Vivalt, Eva (2017a). "How Much Can We Generalize from Impact Evaluations?", working paper.

Vivalt, Eva (2017b). "The Trajectory of Specification Searching Across Disciplines and Methods", working paper.

World Bank (2015). "Country and Lending Groups", http://data.worldbank.org/about/country-and-lending-groups.

USAID (2011). "Evaluation: Learning from Experience", USAID Evaluation Policy,

Washington, DC.

Young, Alwyn (2016). "Channelling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results", working paper.

Appendices

A Guide to Online Appendices

Appendix B provides illustrations of the elicitation mechanism and other questions used in the experiment, and Appendix C provides additional results.

Having to describe an experiment and data from twenty different meta-analyses and systematic reviews, however, I must rely in part on online appendices. The following are available at http://www.evavivalt.com/appendices-learning:

D) Derivation of the equations governing updating in a mixed model.

E) Excerpt from AidGrade's Process Description (2013).

F) The search terms and inclusion criteria for each topic.

G) The references for each topic.

H) The coding manual.

B Experimental Details

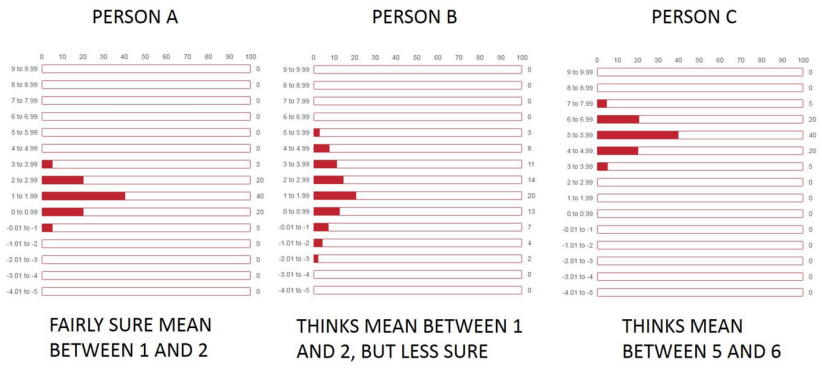
The following diagrams are excerpted from the survey.

Figure 7: Sample Screening Question

Several simple screening questions were used. After this question, respondents were presented with data and then asked to provide another estimate.

Figure 8: Understanding Check

You will also be asked to provide your best estimate of what the true program impact is using a slider like the one below. The number of points you assign to each row will directly correspond to how likely you think the true impact was to fall within that range. Take a look at the following examples.



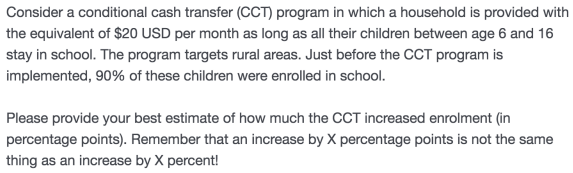
For instance, person A and B both suggest that the impact of a program is most likely to be in the range of 1 – 2 percentage points, while person C thinks the most likely range is between 3 – 4 percentage points.

Person A is much more confident that the program had an effect around 1 or 2 percentage points than person B since person A assigns lower weights to numbers outside of this range compared to person B.

Do these examples make sense to you?

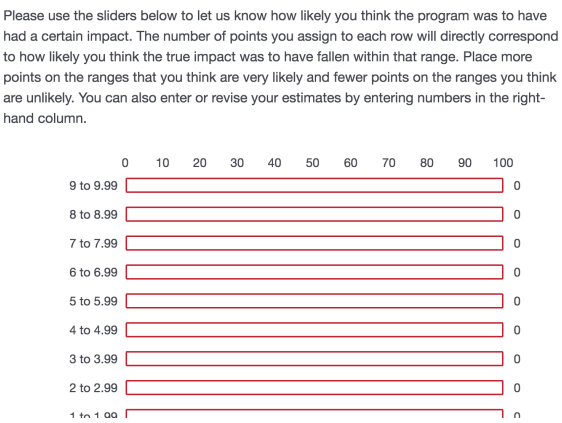
Respondents were walked through several examples of how they might distribute weights to different bins. MTurk respondents were provided with the accompanying written text describing each picture, while policymakers were provided with this information orally.

Figure 9: Sample Program Description



Respondents were provided with a short description of a conditional cash transfer program and a school meals program, then asked to provide their best guess as to the effect of the program.

Figure 10: Assigning Likelihoods



Respondents were then asked to use slider bars to place weights on the probability of different outcomes.

C Additional Results

Table 8: Descriptive Statistics: Narrowly Defined Outcomes

Intervention	Outcome	# Neg sig papers	# Insig papers	# Pos sig papers	# Papers
Conditional cash transfers	Attendance rate	0	6	9	15
Conditional cash transfers	Enrollment rate	0	7	29	36
Conditional cash transfers	Gave birth at healthcare facility	0	2	1	3
Conditional cash transfers	Height	0	1	1	2
Conditional cash transfers	Height-for-age	0	6	1	7
Conditional cash transfers	Labor force participation	1	12	5	18
Conditional cash transfers	Labor hours	0	3	4	7
Conditional cash transfers	Pregnancy rate	1	1	1	3
Conditional cash transfers	Probability unpaid work	1	0	4	5
Conditional cash transfers	Retention rate	0	3	2	5
Conditional cash transfers	Skilled attendant at delivery	0	3	0	3
Conditional cash transfers	Test scores	1	2	2	5
Conditional cash transfers	Unpaid labor hours	3	2	0	5
Conditional cash transfers	Weight-for-age	0	2	0	2
Conditional cash transfers	Weight-for-height	0	1	1	2
HIV/AIDS Education	Contracted STD	0	2	1	3
HIV/AIDS Education	Has multiple sex partners	0	2	2	4
HIV/AIDS Education	Pregnancy rate	0	1	3	4
HIV/AIDS Education	Probability sexually active	0	2	1	3
HIV/AIDS Education	Used contraceptives	0	2	8	10
Unconditional cash transfers	Enrollment rate	0	5	8	13
Unconditional cash transfers	Test scores	0	1	1	2
Unconditional cash transfers	Weight-for-height	0	2	0	2
Insecticide-treated bed nets	Malaria	0	4	14	18
Contract teachers	Test scores	0	1	2	3
Deworming	Attendance rate	0	1	1	2
Deworming	Birthweight	0	2	0	2
Deworming	Diarrhea incidence	0	1	1	2
Deworming	Height	3	10	3	16
Deworming	Height-for-age	1	9	4	14
Deworming	Hemoglobin	0	13	1	14
Deworming	Malformations	0	2	0	2
Deworming	Mid-upper arm circumference	2	0	5	7
Deworming	Test scores	0	0	2	2
Deworming	Weight	3	8	6	17
Deworming	Weight-for-age	1	6	5	12
Deworming	Weight-for-height	2	7	2	11

Financial literacy	Has savings	0	4	1	5
Financial literacy	Has taken loan	0	4	0	4
Financial literacy	Savings	0	2	3	5
Improved stoves	Chest pain	0	0	2	2
Improved stoves	Cough incidence	0	0	2	2
Improved stoves	Difficulty breathing	0	0	2	2
Improved stoves	Excessive nasal secretion	0	1	1	2
Irrigation	Consumption	0	1	1	2
Irrigation	Total income	0	1	1	2
Microfinance	Assets	0	3	1	4
Microfinance	Consumption	0	2	0	2
Microfinance	Probability of owning business	0	1	1	2
Microfinance	Profits	1	3	1	5
Microfinance	Savings	0	3	0	3
Microfinance	Total income	0	3	2	5
Micro health insurance	Enrollment rate	0	1	1	2
Micro health insurance	Household health expenditures	0	1	1	2
Micro health insurance	Probability of inpatient visit	0	2	0	2
Micro health insurance	Probability of outpatient visit	0	2	0	2
Micronutrient supplementation	Birthweight	0	4	3	7
Micronutrient supplementation	Body mass index	0	1	4	5
Micronutrient supplementation	Cough incidence	0	1	1	2
Micronutrient supplementation	Cough prevalence	0	2	1	3
Micronutrient supplementation	Diarrhea incidence	0	3	10	13
Micronutrient supplementation	Diarrhea prevalence	0	5	8	13
Micronutrient supplementation	Fever prevalence	0	2	1	3
Micronutrient supplementation	Height	3	19	7	29
Micronutrient supplementation	Height-for-age	4	21	8	33
Micronutrient supplementation	Hemoglobin	6	11	20	37
Micronutrient supplementation	Malaria	0	0	3	3
Micronutrient supplementation	Mid-upper arm circumference	2	8	7	17
Micronutrient supplementation	Mortality	1	10	1	12
Micronutrient supplementation	Perinatal death	0	5	1	6
Micronutrient supplementation	Prevalence of anemia	0	0	13	13
Micronutrient supplementation	Stillbirth	0	0	4	4
Micronutrient supplementation	Stunted	0	0	3	3
Micronutrient supplementation	Test scores	1	2	6	9
Micronutrient supplementation	Triceps skinfold measurement	1	0	1	2
Micronutrient supplementation	Weight	1	17	13	31

Micronutrient supplementation	Weight-for-age	1	20	10	31
Micronutrient supplementation	Weight-for-height	0	18	8	26
Mobile phone-based reminders	Appointment attendance rate	0	0	3	3
Mobile phone-based reminders	Treatment adherence	0	2	3	5
Performance pay	Test scores	0	2	1	3
Rural electrification	Enrollment rate	0	1	2	3
Rural electrification	Study time	0	1	2	3
Rural electrification	Total income	0	2	0	2
Safe water storage	Diarrhea incidence	0	0	2	2
Scholarships	Attendance rate	0	1	1	2
Scholarships	Enrollment rate	0	2	1	3
Scholarships	Test scores	0	2	0	2
School meals	Enrollment rate	0	3	0	3
School meals	Height-for-age	0	2	0	2
School meals	Test scores	0	2	1	3
Water treatment	Diarrhea incidence	0	1	5	6
Water treatment	Diarrhea prevalence	0	3	7	10
Water treatment	Dysentery incidence	0	1	2	3
Women's empowerment programs	Savings	0	1	1	2
Women's empowerment programs	Total income	0	0	2	2
Average		0.4	3.6	3.3	7.3

Table 9: Marginal Benefits of an Additional Study, $\theta^* = 0.5\mu$, For Each Intervention-Outcome

Intervention	Outcome	$B_{1,2}$	$B_{5,6}$	$B_{10,11}$	N
SMS Reminders	Treatment adherence	-0.0036			4
Scholarships	Enrollment rate	-0.0030			4
SMS Reminders	Appointment attendance rate	-0.0010			2
Micronutrients	Weight	-0.0004	0.0003	0.0001	35
Microfinance	Assets	-0.0001			3
Micronutrients	Weight-for-height	0.0000	0.0003	0.0002	25
Conditional Cash Transfers	Attendance rate	0.0000	0.0077	0.0000	14
Conditional Cash Transfers	Labor force participation	0.0000	-0.0008	0.0008	16
Bed Nets	Malaria	0.0000	0.0000		8
Contract Teachers	Test scores	0.0000			2
Micronutrients	Prevalence of anemia	0.0000	0.0016	0.0128	14
Rural Electrification	Enrollment rate	0.0000			2
Unconditional Cash Transfers	Enrollment rate	0.0000	0.0071		10
Micronutrients	Height-for-age	0.0001	0.0012	0.0005	35
Micronutrients	Mortality rate	0.0001	0.0000	0.0000	11
Micronutrients	Weight-for-age	0.0001	0.0001	0.0000	33
Conditional Cash Transfers	Enrollment rate	0.0001	0.0000	0.0038	36
Conditional Cash Transfers	Height-for-age	0.0001	0.0006		6
Micronutrients	Stillbirths	0.0003			3
Deworming	Hemoglobin	0.0004	-0.0001	0.0007	14
Micronutrients	Height	0.0005	0.0002	0.0002	31
Microfinance	Profits	0.0006			4
Micronutrients	Cough prevalence	0.0006			2
Micronutrients	Mid-upper arm circumference	0.0006	0.0005	0.0006	17
Deworming	Weight-for-height	0.0007	0.0042		10
Micronutrients	Diarrhea incidence	0.0010	0.0069		10
Deworming	Height	0.0012	0.0015	0.0006	16
Micronutrients	Hemoglobin	0.0013	0.0006	0.0076	45
Financial Literacy	Savings	0.0014			4
Microfinance	Total income	0.0016			4
Deworming	Weight	0.0019	0.0011	0.0016	17

Microfinance	Savings	0.0020			2
HIV/AIDS Education	Used contraceptives	0.0026	0.0021		9
Micronutrients	Birthweight	0.0028	0.0046		6
Micronutrients	Perinatal deaths	0.0034			5
Conditional Cash Transfers	Probability unpaid work	0.0035			4
Deworming	Height-for-age	0.0042	0.0019	0.0051	13
Deworming	Weight-for-age	0.0051	0.0027	0.0053	11
Micronutrients	Fever prevalence	0.0056			4
Micronutrients	Body mass index	0.0062			4
Micronutrients	Test scores	0.0064	0.0023		9
Conditional Cash Transfers	Unpaid labor	0.0066			4
School Meals	Test scores	0.0071			2
Water Treatment	Diarrhea prevalence	0.0088			5
Conditional Cash Transfers	Test scores	0.0092			4
Micronutrients	Diarrhea prevalence	0.0094			5
School Meals	Enrollment rate	0.0126			2
Micronutrients	Stunted	0.0159			4
Performance Pay	Test scores	0.0200			2
Deworming	Mid-upper arm circumference	0.0351	0.0191		6
Rural Electrification	Study time	0.1586			2
Average		0.0065	0.0025	0.0023	11
Median		0.0010	0.0011	0.0006	6

Table 10: Marginal Benefits of an Additional Study, $\theta^* = 0.9\mu$, For Each Intervention-Outcome

Intervention	Outcome	$B_{1,2}$	$B_{5,6}$	$B_{10,11}$	N
SMS Reminders	Appointment attendance rate	-0.0054			2
Performance Pay	Test scores	-0.0037			2
Bed Nets	Malaria	-0.0005	0.0014		8
Deworming	Weight	-0.0003	0.0002	0.0003	17
Micronutrients	Perinatal deaths	-0.0003			5
Micronutrients	Stillbirths	-0.0002			3
Conditional Cash Transfers	Labor force participation	-0.0002	0.0000	0.0009	16
School Meals	Enrollment rate	-0.0001			2
Micronutrients	Birthweight	-0.0001	-0.0009		6
Micronutrients	Fever prevalence	0.0000			4
Unconditional Cash Transfers	Enrollment rate	0.0000	0.0000		10
Rural Electrification	Enrollment rate	0.0000			2
Micronutrients	Cough prevalence	0.0000			2
Micronutrients	Mortality rate	0.0000	0.0000	0.0000	11
Micronutrients	Weight-for-age	0.0000	0.0001	0.0000	33
Micronutrients	Weight-for-height	0.0000	0.0001	0.0001	25
Water Treatment	Diarrhea prevalence	0.0000			5
Micronutrients	Height	0.0000	0.0000	0.0000	31
Micronutrients	Height-for-age	0.0001	0.0000	0.0001	35
Conditional Cash Transfers	Height-for-age	0.0001	0.0000		6
Deworming	Hemoglobin	0.0001	-0.0002	0.0000	14
Microfinance	Profits	0.0001			4
Deworming	Height	0.0001	0.0001	0.0002	16
Microfinance	Assets	0.0002			3
Micronutrients	Weight	0.0002	0.0001	0.0000	35
Deworming	Weight-for-height	0.0002	0.0001		10
Financial Literacy	Savings	0.0003			4
Micronutrients	Mid-upper arm circumference	0.0003	0.0000	0.0000	17
Micronutrients	Hemoglobin	0.0003	-0.0002	-0.0001	45
Deworming	Height-for-age	0.0004	0.0016	0.0012	13
Microfinance	Savings	0.0004			2

Conditional Cash Transfers	Attendance rate	0.0004	0.0009	0.0019	14
Micronutrients	Prevalence of anemia	0.0005	0.0004	0.0012	14
Conditional Cash Transfers	Probability unpaid work	0.0005			4
Microfinance	Total income	0.0006			4
Deworming	Mid-upper arm circumference	0.0006	0.0069		6
Conditional Cash Transfers	Unpaid labor	0.0006			4
Micronutrients	Diarrhea incidence	0.0008	0.0010		10
Conditional Cash Transfers	Test scores	0.0008			4
Micronutrients	Stunted	0.0009			4
Deworming	Weight-for-age	0.0011	0.0009	-0.0001	11
HIV/AIDS Education	Used contraceptives	0.0011	0.0001		9
Micronutrients	Test scores	0.0014	0.0007		9
Micronutrients	Diarrhea prevalence	0.0014			5
SMS Reminders	Treatment adherence	0.0015			4
School Meals	Test scores	0.0016			2
Conditional Cash Transfers	Enrollment rate	0.0016	0.0000	0.0000	36
Rural Electrification	Study time	0.0022			2
Scholarships	Enrollment rate	0.0042			4
Micronutrients	Body mass index	0.0049			4
Contract Teachers	Test scores	0.0061			2
Average		0.0005	0.0005	0.0003	11
Median		0.0002	0.0001	0.0000	6

Table 11: Marginal Benefits of an Additional Study, N $\geqslant$ 10

	$B_{1,2}$	$B_{5,6}$	$B_{10,11}$
$\theta^* = 0.5\mu$			
20th percentile	0.0000	0.0000	0.0000
40th percentile	0.0001	0.0002	0.0003
60th percentile	0.0006	0.0007	0.0007
80th percentile	0.0019	0.0047	0.0027
$\theta^* = 0.9\mu$			
20th percentile	0.0000	0.0000	0.0000
40th percentile	0.0001	0.0001	0.0001
60th percentile	0.0002	0.0002	0.0002
80th percentile	0.0005	0.0006	0.0004

This table shows the calculated marginal benefits, $B_{i,j}$, of moving from study i to study j, assuming that $\theta^* = 0.5\mu$ or $\theta^* = 0.9\mu$, for those intervention-outcome combinations with at least 10 studies. To generate this figure, we form each possible order of studies' results within each intervention-outcome, calculate whether, in each case, the jth study would be pivotal and in which direction, and then take the expected value of the benefits across all the different possible ways to move from i to j studies within that intervention-outcome. All benefits are in terms of effect sizes and all $B_{i,j}$ calculations are done on whichever intervention-outcomes have at least j studies, meaning that the columns are not strictly comparable to each other as different intervention-outcome combinations could be included in each. As a point of reference, the typical effect size of a study is 0.12.