# Weighing the Evidence: Which Studies Count?

Eva Vivalt* Aidan Coville† Sampada KC‡

April 30, 2021

**Abstract**

We present results from two experiments run at World Bank and Inter-American Development Bank workshops on how policy-makers, practitioners and researchers weigh evidence and seek information from impact evaluations. We find that policy-makers care more about attributes of studies associated with external validity than internal validity, while for researchers the reverse is true. These preferences can yield large differences in the estimated effects of pursued policies: policy-makers indicated a willingness to accept a program that had a 6.3 percentage point smaller effect on enrollment rates if it were recommended by a local expert, larger than the effects of most programs.

*Department of Economics, University of Toronto, eva.vivalt@utoronto.ca

†Development Impact Evaluation group, World Bank, acoville@worldbank.org

‡Institutions and Political Inequality, Berlin Social Science Center (WZB), sampada.kc@wzb.eu.

# 1 Introduction

The basic motivation behind evidence-based policy-making is simple: rigorous evidence on what works best could help policy-makers make more informed decisions that improve policy effectiveness. This could take the form of, for example, scale up or adaptation of programs based on evidence on the project of interest or similar projects elsewhere. The first constraint to effective use of evidence in policy is the availability of relevant research. The last twenty years have seen a surge of new studies in developing countries to help tackle this constraint. As the research base grows for a particular topic, this opens the opportunity for selective use of evidence - focusing on research that is more relevant for the policy questions at hand. In many cases this selective use of evidence may be well-justified – for example, limiting attention to higher-quality, or more contextually relevant studies. However, since selective use of evidence may result in different guidance for policy, a natural question arises: how do policymakers select and assess evidence to focus on when making decisions? Does this differ in systematic ways from how others, such as researchers, may assess the same pool of evidence? Finally, could this selection process influence beliefs and decisions?

To explore these questions, we leverage a unique opportunity to run a set of experiments with policy-makers, practitioners and researchers invited to World Bank (WB) and Inter-American Development Bank (IDB) impact evaluation workshops. The workshops are designed as "matchmaking" events between those involved in certain development programs and researchers. Our focus on impact evaluation workshop participants is intentional: higher-level policy-makers will often not have time to read academic papers themselves and might instead rely on briefings from advisors. Workshop attendees are also demonstrably interested in impact evaluation and thus may be a particularly relevant sample. Workshop participants include program officers in government agencies; monitoring and evaluation specialists within government agencies; World Bank or IDB operational staff; and other aid agency operational staff such as technical advisors at USAID or the FCDO {footnoteUSAID and the FCDO are the bilateral development aid agencies of the United States and United Kingdom, respectively.. The sample also includes a group of researchers, both from academic institutions as well as some international organizations.

We run two experiments at these workshops. First, we leverage a discrete choice

experiment to consider how policy-makers and practitioners *weigh* the value of different kinds of evidence they are presented with in selecting *programs*. The kinds of evidence they consider comprise: evidence from RCTs, quasi-experimental or observational studies; results with greater or less precision; and evidence from different locations. We also compare how they weigh these impact evaluation results relative to advice from a *local expert*. This latter type of evidence is frequently used due to the importance of context and the fact that local impact evaluation results are often unavailable, but our paper is the first to directly consider it. We find that in comparison to researchers, who place a high weight on the methods of impact evaluations of the programs, policy-makers value contextual factors such as whether a local expert recommended the program and put relatively less weight on research methods. In other words, policy-makers value factors one might associate with external validity over factors one might associate with internal validity.

We also leverage a second set of discrete choice experiments to consider how policy-makers, practitioners and researchers *seek* information from *studies*. Weighing information when selecting a program and seeking information from studies are related but distinct. For example, policy-makers could seek information from an impact evaluation that estimated that a program had no effect in order to learn what went wrong, however, all else equal, they would not want to choose a program estimated to have no effects when weighing which program to select. In this second experiment, apart from method and location, we consider implementing organization and sample size, both of which have been shown to be important predictors of treatment effects (Bold et al., 2018; Cameron et al. 2019; Ioannidis, Stanley and Doucouliagos, 2017; Vivalt, 2020). Again, we find that researchers care more about study attributes associated with internal validity while policy-makers and practitioners care more about factors associated with external validity. Policy-makers are also more likely to select studies with larger estimated treatment effects, a finding which may help to explain why policy-makers tend to predict that programs will have larger effects than researchers predict they will have (de Andrade et al., 2014; Hirschleifer et al., 2014; Casey et al., 2019; Vivalt and Coville, 2020).

However, policy-makers and researchers weigh results more similarly than researchers predict. Prior to obtaining final results, we ran a forecasting exercise on the Social Science Prediction Platform. Here, researchers participated in the experiment and forecast policy-makers', practitioners' and other researchers' responses.

Researchers generally over-estimated how much weight policy-makers would place on each attribute but, accounting for this, they anticipated policy-makers would place relatively less weight on factors associated with internal validity than they did in practice. Interestingly, researchers also under-estimated the extent to which researchers value local expertise. In short, the groups behave more similarly than expected.

It can be challenging to find a setting in which one can run experiments on policy-makers, so evidence about how policy-makers weigh impact evaluation results is sparse. Banuri et al. (2015) demonstrate that policy practitioners may be subject to behavioral biases. Nellis et al. (2019) investigate how development practitioners learn from meta-analysis results as opposed to individual studies. Hjort et al. (2019) leverage a sample of mayors to run two experiments, one considering the impact of providing information on the efficacy of tax reminder letters, along with template letters, on implementation, and the other looking at individuals' willingness-to-pay for information from impact evaluations. This paper is perhaps the closest to ours, though ours diverges in several key respects. First, knowing that policy advice and evidence from impact evaluations can conflict, we examine how policy-makers weigh impact evaluation results compared to advice from *local experts*. Second, we ask policy-makers to select not just *studies* but *programs*. In doing so, we deliberately present policy-makers with the results of the impact evaluations, which allows us to examine how they trade off estimated treatment effects with the source of information. This allows us to say, for example, that policy-makers would accept a program that was not recommended by a local expert over one that was only if the former had at least a 6.3 percentage point higher estimated impact on enrollment rates, and they would prefer a program evaluated in a different region over one evaluated in their country only if the program evaluated in a different region had at least a 4.5 percentage point higher estimated impact. These estimated impacts are often larger than the typical effects of popular programs that improve enrollment rates.[1] Capturing willingness-to-pay in this way thus enables us to consider the potential magnitude of these trade-offs in terms of the underlying public good at stake. Finally, we consider a wider range of study characteristics associated with internal and external validity, compare our results to what researchers currently expect, and connect these measures with a large data set of real study results.

[1]For example, in our data, the median treatment effect of 36 conditional cash transfer programs on enrollment rates was 5.1 percentage points.

Our paper also relates to the growing evidence on how the same intervention may have different effects across contexts (Bold et al., 2018; Cameron et al., 2019; Vivalt, 2020). Given that we observe differences in how researchers and policymakers weigh factors associated with internal and external validity, a natural question is: who is right? We cannot fully answer this question but present some back-of-the-envelope calculations by applying our willingness to pay estimates to a larger data set of 635 impact evaluation results to predict each group's preferred subset of studies.

The rest of the paper proceeds as follows. First, we discuss the data used in each experiment. Then we describe and present results from the set of discrete choice experiments on how policy-makers, practitioners and researchers weigh evidence when selecting programs. We turn to the set of discrete choice experiments on how policy-makers, practitioners and researchers seek information from impact evaluations. Finally, we discuss the implications.

# 2 Data

We conducted surveys with policy-makers at different Inter-American Development Bank and World Bank workshops. Table 1 provides a list. Each of the samples is described in greater detail below.

### 2.0.1 World Bank Sample

We surveyed participants at workshops organized in Mexico City (May 2016), Nairobi (June 2016), Athens (September 2019) and Marrakesh (December 2019). Workshop attendees comprised policy-makers, practitioners, and researchers. The workshops were each approximately one week long and were designed as "matchmaking" events between government staff and researchers. Government counterparts were paired with researchers and required 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 the FCDO; 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 "policy-makers"; international organization operational staff and NGO

Table 1: Response Rate at Workshops

| Institution | Location | Year | Eligible Attendees | Surveyed | Response Rate |
|---|---|---|---|---|---|
| *Experiment 1: Weighing Evidence* | | | | | |
| IDB | Washington, D.C. | 2018 | 49 | 18 (18) | 0.37 (0.37) |
| World Bank | Athens, Greece | 2019 | 39 | 38 | 0.97 |
| World Bank | Marrakesh, Morocco | 2019 | 41 | 33 | 0.80 |
| *Experiment 2: Seeking Evidence* | | | | | |
| World Bank | Mexico City, Mexico | 2016 | 195 | 43 | 0.22 |
| World Bank | Nairobi, Kenya | 2016 | 72 | 49 | 0.68 |
| IDB | Washington, D.C. | 2016 | 75 | 37 (37) | 0.49 (0.49) |
| IDB | Washington, D.C. | 2017 | 62 | 31 (17) | 0.50 (0.27) |

The IDB rows include responses from the "pre" period and, in parentheses, the "post" period. Experiment #1 excludes researchers from both the eligible and response counts, as too few attended to be considered. This excludes 2 researchers' responses from the IDB workshop, 12 from the World Bank workshop in Athens and 2 from the World Bank workshop in Marrakesh. In total, these 14 individuals made only 58 selections in the discrete choice experiment. Researchers were also not included in the numbers for those "eligible" for Experiment 1, except for the case of the IDB where the 49 "eligible" attendees excludes only those 2 known to be researchers and the true number of eligible participants is likely lower.

or private sector employees 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.

Respondents were asked to take these experiments as part of the workshops. No incentives were provided. Policy-makers, practitioners and researchers were not restricted to go to workshops in their geographic area, and the workshops attracted participants from around the world. For example, someone from Nepal might attend the workshop in Mexico. Attendees at the workshops in Mexico and Kenya were asked to participate in the discrete choice experiment on seeking information and attendees at the workshops in Greece and Morocco were asked to participate in the discrete choice experiment on weighing information.

#### 2.0.2 IDB Sample

We also surveyed participants at three separate workshops organized in June 2016, June 2017, and May 2018 at the IDB headquarters in Washington, DC. These workshops similarly brought together policy-makers and practitioners for a week each, to provide training on impact evaluation methods. The surveys were distributed by the workshop organizers, who sent an email with a link to the survey before the workshops began and after they ended. Our results focus on the responses obtained before the workshop began, as they may be closer to the typical preferences of policy-makers and practitioners. Participation was encouraged but voluntary, and no incentives were provided. One difference from the World Bank workshops is that researchers were not asked to take these surveys. Attendees at the 2016 and 2017 workshops were asked to participate in the discrete choice experiment on seeking information and attendees at the 2018 workshop were asked to participate in the discrete choice experiment on weighing information.

# 3 Experiment 1: Weighing Evidence

## 3.1 Method

In the first experiment, participants were asked which one of two conditional cash transfer programs they would recommend for implementation. The programs were simply labeled as *Program A* and *Program B* and were intended to raise school

Table 2: Attributes and Levels used for IDB 2018, Athens, and Marrakesh Sample

| Attributes | Levels |
|---|---|
| Method | Experimental, Quasi-experimental, Observational |
| Location | Different country, Same country, Different country in the same region |
| Impact | 0, +5, +10 percentage points |
| Confidence Interval | $+/-1$, $+/-10$ percentage points |
| Recommended | Yes, No |

enrollment. Each program had an impact evaluation associated with it, and the impact evaluation differed by the method used (experimental, quasi-experimental, observational); location (same country, different country in the same region, different country in a different region); precision (a confidence interval of +/- 1 percentage point or +/- 10 percentage points); whether a local expert recommended it; and the effect the study found (an increase in enrollment rates by 0, 5, or 10 percentage points). These attributes are summarized in Table 2. Appendix figure B1 illustrates an example of a choice scenario faced by participants.

We chose these attributes due to their relevance for evidence-based policy-making. "Method" and "precision" are important when considering the internal validity of a study's results. "Location" is often used as an indicator of external validity, though Vivalt (2020) finds that program implementer may be more informative. Whether or not a program is recommended by a local expert may also provide further evidence relevant to their context. An estimate of the program impact was included to help gauge individuals' willingness-to-pay for different factors - *i.e.* how much of a decrease in estimated treatment effect they would be willing to accept in exchange for better quality evidence.

Given the number of attributes and their levels, a full factorial design would yield an impractically large number of choice sets. Instead, we used a fractional factorial design with questions grouped into blocks.[2] A block consisted of six choice sets of two alternatives each. We randomized the blocks across respondents and the questions within blocks. Individual respondents saw one block each at the World Bank workshops in Athens and Marrakesh and two blocks at the IDB. We also constructed

[2]Using the *dcreate* package in Stata for a D-efficient design.

an indicator variable for whether the result shown was significant. While this variable was not explicitly shown to participants, it could be discerned from the provided estimated impact and confidence interval. Results are limited to policy-makers and practitioners, as few researchers attended these particular workshops.

## 3.2 Results

We first analyzed the data using a conditional logit, in which the dependent variable takes the value of 1 for the chosen alternative in each choice set and 0 otherwise. Table 3 presents results. Policy-makers preferred programs with larger estimated treatment effects or programs that came recommended by a local expert. Practitioners preferred programs with larger, more precisely estimated impact evaluation results as well as results from the same country as the target program and results from RCTs. In an alternative specification (Appendix Table B1), we use a mixed model treating the coefficient for *impact* as fixed and assuming a normal distribution for the remaining variables.[3] Results are comparable. If we create a dummy variable indicating whether a result is significant or not, and include it in the regression, this appears to have driven the preference towards results with a small confidence interval (Appendix Table B2). Results appear largely comparable across the World Bank and IDB pre-workshop samples.[4] Results that are significant for a subgroup in one sample will not always be significant in the same subgroup in the other sample, but this may be due to the smaller sample sizes in these disaggregated analyses, and the magnitudes of the coefficients are mostly aligned. Our preferred specification pools the samples for increased power.

What do these results mean in terms of how policy-makers may be willing to make trade-offs between programs supported by different types of evidence? In Table 4, we present estimates of participants' willingness-to-pay in terms of estimated impact. In this table, we can see for example that policy-makers would be willing to accept a program with a 6.3 percentage point lower estimated impact if it came recommended

[3]We use 1,000 Halton draws to estimate the mixed logit model.

[4]Table B3 shows results separately for those who took both rounds of the IDB experiment and those who took only one round. These results are also split by whether they were obtained "pre" or "post" workshop. Among those who took both rounds of the survey, an insignificant weight was placed on RCTs in the "pre" workshop survey and a significant weight was placed on RCTs in the "post" period. Recalling that the IDB sample consisted of policy-makers and practitioners, this makes intuitive sense: they may have been less familiar with impact evaluation methods *ex ante*.

by a local expert. They would likewise be willing to accept a program with 4.5 percentage point lower estimated effects so long as that estimate came from the same country as the target country. These results imply that unless research is seen by policy-makers as valid in their target setting, policy-makers are likely to choose alternative programs that may have lower estimated treatment effects but be from a better-fitting context.

How do these results compare with researcher forecasts? We gathered forecasts from policy-makers, practitioners and researchers through the Social Science Prediction Platform July 8, 2020 - August 17, 2020, as one of the first studies made available on the platform for others to forecast. Participants were recruited both by targeted emails and via a survey link shared on Twitter. The main focus was on gathering forecasts from researchers. 159 researchers responded to the survey, including 21 (or 50%) of those invited by personalized email. Importantly, as part of the forecasting exercise, researchers were first asked to participate in the same discrete choice experiment as the policy-makers and practitioners did, answering one randomly-selected block of six questions. This gave them familiarity with the questions the policy-makers and practitioners saw. After this exercise, they were asked to forecast policy-makers' and practitioners' choices, separately. Finally, they predicted how researchers taking the forecasting survey weighed the different attributes when they completed the exercise. Details of the forecasting exercise are provided in an appendix.

Figure 1 shows a scatter plot of mean researcher forecasts with results. The first thing to notice from this figure is that almost all forecasts were larger in magnitude than our results, as illustrated by the difference between forecasts of researchers' choices and the 45 degree line. Still, researcher forecasts of policy-makers' responses were 0.75 correlated with the true results; their forecasts of practitioners' responses were essentially uncorrelated (-0.01) with the results; and their forecasts of researchers' responses in the forecasting survey had a correlation coefficient of 0.51.

We need to be cautious when comparing across researcher and policy-maker responses since they were collected in different settings - the former through the Social Science Prediction platform, and the latter in the workshops. Instead of relying on the point estimates, we consider the rank order of forecasts and results. Looking at the figure and reading it from the top down, researchers thought researchers would put the most weight on method, followed by impact, having small confidence intervals, and whether it was done in the same country, in that order, putting the least

weight on being recommended by a local expert. In reality, reading the figure from right to left, researchers put the *most* weight on impact and the *third*-most weight on being recommended by a local expert, relatively more than they thought. Researchers thought policy-makers would put the most weight on location, followed by impact, and whether it was recommended by a local expert, and less weight on method or confidence interval. In reality, policy-makers put the most weight on impact and whether it was recommended by a local expert, followed by some small weight on being an RCT. The main takeaway from these forecasts is that while policy-makers and researchers weigh evidence differently, they may be more similar in their relative weighting of study attributes than researchers believe.

Table 3: Weighing Evidence from Research Results vs. Local Experts

| | *Pooled* | | *World Bank* | | *IDB* | |
|---|---|---|---|---|---|---|
| | Policy-maker | Practitioner | Policy-maker | Practitioner | Policy-maker | Practitioner |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Impact | 1.068*** | 1.107*** | 1.056** | 1.094*** | 1.081** | 1.134*** |
| | (0.023) | (0.024) | (0.029) | (0.029) | (0.039) | (0.041) |
| Quasi-Experimental | 0.935 | 1.385 | 0.863 | 1.841** | 0.986 | 0.913 |
| | (0.189) | (0.275) | (0.224) | (0.476) | (0.343) | (0.300) |
| Experimental | 1.055 | 1.674*** | 1.030 | 1.895** | 1.055 | 1.461 |
| | (0.204) | (0.329) | (0.254) | (0.503) | (0.358) | (0.432) |
| Different country, same region | 0.994 | 1.309 | 1.192 | 1.468 | 0.727 | 1.061 |
| | (0.191) | (0.255) | (0.293) | (0.377) | (0.238) | (0.337) |
| Same country | 1.339 | 1.924*** | 1.240 | 2.080*** | 1.519 | 1.751* |
| | (0.261) | (0.376) | (0.313) | (0.529) | (0.484) | (0.570) |
| Recommended | 1.513*** | 1.183 | 1.348 | 1.066 | 1.761** | 1.397 |
| | (0.223) | (0.163) | (0.258) | (0.191) | (0.427) | (0.309) |
| Small CI | 1.206 | 1.483*** | 1.147 | 1.441** | 1.200 | 1.581* |
| | (0.169) | (0.205) | (0.211) | (0.261) | (0.271) | (0.373) |
| Observations | 239 | 267 | 143 | 156 | 96 | 111 |

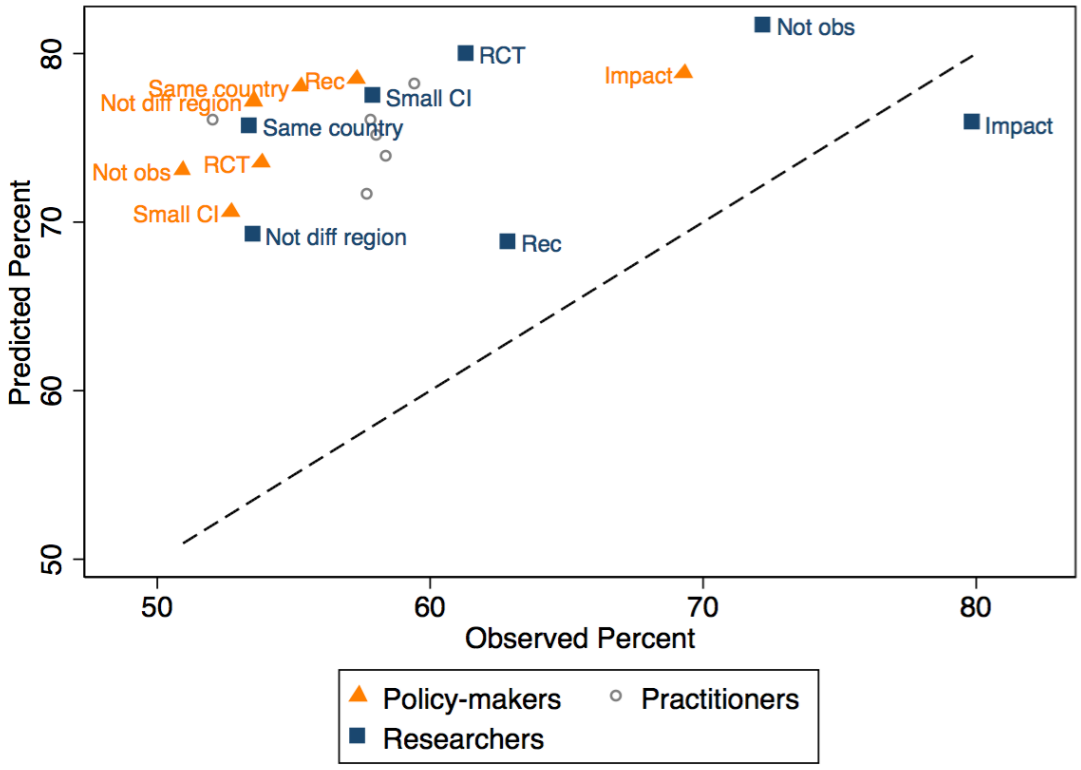
This table reports the results of conditional logit regressions on which program was selected, using odds ratios. The omitted categories are "Observational" and "Different region". The number of observations represents the total number of choices made across individuals. The IDB results use only the pre-workshop sample. Standard errors are provided in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 4: Willingness to Pay (in Terms of Estimated Impact)

| | *Pooled* | | *World Bank* | | *IDB* | |
|---|---|---|---|---|---|---|
| | Policy-maker | Practitioner | Policy-maker | Practitioner | Policy-maker | Practitioner |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Quasi-Experimental | 1.03 | -3.21 | 2.71 | -6.77 | 0.18 | 0.72 |
| | (3.11) | (1.99) | (5.05) | (3.31) | (4.46) | (2.65) |
| Experimental | -0.82 | -5.08 | -0.54 | -7.09 | -0.69 | -3.01 |
| | (2.98) | (2.14) | (4.56) | (3.57) | (4.36) | (2.34) |
| Different country, | 0.09 | -2.66 | -3.23 | -4.26 | 4.09 | -0.47 |
| same region | (2.93) | (2.00) | (4.91) | (3.09) | (4.50) | (2.53) |
| Same country | -4.47 | -6.46 | -3.96 | -8.12 | -5.35 | -4.45 |
| | (3.08) | (2.13) | (4.76) | (3.43) | (4.16) | (2.65) |
| Small CI | -2.86 | -3.89 | -2.53 | -4.05 | -2.33 | -3.64 |
| | (2.25) | (1.46) | (3.49) | (2.24) | (2.99) | (1.87) |
| Recommended | -6.33 | -1.65 | -5.49 | -0.71 | -7.24 | -2.66 |
| | (2.57) | (1.36) | (3.72) | (1.98) | (4.01) | (1.80) |
| Observations | 239 | 267 | 143 | 156 | 96 | 111 |

This table reports the results of conditional logit regressions in terms of the implied willingness-to-pay for programs with certain attributes. For example, in the pooled sample, policy-makers would only be willing to accept a program that was not recommended over one that was if the program that was not recommended had a 6.33 percentage point higher estimated impact on enrollment rates. The number of observations represents the total number of choices made across individuals. The IDB results use only the pre-workshop sample. Standard errors are provided in parentheses.

Figure 1: Researcher Forecasts




This figure shows the mean forecast researchers made for each subgroup-attribute combination, as plotted against the values that were observed in practice. To simplify the forecasting exercise, forecasters were asked to predict the percent of time that respondents in each subgroup would select a certain option. Attributes with more than two levels were collapsed to two in the forecasting survey (e.g., RCT and non-RCT; observational and non-observational). Forecasters were reminded that even if a respondent preferred a certain level of an attribute (e.g., RCTs to non-RCTs), that would not translate to their choosing the option with that level of the attribute 100% of the time due to random variation in the other attributes, and a value of 50% would indicate indifference regarding that attribute. "Rec" represents "Recommended", "Not obs" being not observational (i.e., quasi-experimental or experimental), "Not diff region" being not in a different region (i.e., same country or different country in the same region). The black dashed line represents the 45 degree line of accurate predictions.

# 4 Experiment 2: Seeking Evidence

## 4.1 Method

This experiment, run at World Bank workshops in Nairobi and Mexico City and IDB workshops in Washington, D.C., asked participants to choose which of two *studies* they thought would be most helpful, rather than which of two *programs* they would prefer.[5] While weighing of information and information seeking behavior are clearly related, how individuals seek information could still differ from how they weigh that evidence. For example, someone could wish to learn about an impact evaluation not because they think it did something right but because they want to learn what not to do. What kind of information they seek thus has a more ambiguous interpretation, but since exposure to different information could drive differences in beliefs and policy choice, it is nonetheless important to study.

Respondents saw two blocks of six questions each at all workshops. The studies respondents were offered in this experiment varied by method (experimental, quasi-experimental or observational); location (same country, different country in the same region, different country in a different region); sample size (50, 3,000, or 15,000); program implementer (government or NGO); and the effect the study found (an increase in enrollment rates by 0, 5, or 10 percentage points or a decrease by 5 percentage points). These attributes and levels are summarized in Table 5. More researchers attended the World Bank workshops at which this experiment was run and took the experiment, so we can consider them as an additional subgroup.

## 4.2 Results

Table 6 presents results of a conditional logit. All else equal, participants prefer to seek information from RCTs or quasi-experimental studies, from studies done in the same country or at least the same region as the target program, and from studies with large sample sizes. Interestingly, the coefficient on program implementer is of-

[5] E.g. "Now imagine that you are in charge of developing and implementing a new conditional cash transfer program in your country. You will now be presented with different studies to choose from. Please select one option from each pair of studies that you think will be most useful to help you estimate the likely effect of this program" as opposed to "Now imagine that you need to provide a recommendation to a counterpart agency in your country on which of two programs to implement. A study was done on each program, with the results below. Please select which program you would recommend."

Table 5: Attributes and Levels used for IDB 2016 & 17, Nairobi, and Mexico City Sample

| Attributes | Levels |
|---|---|
| Method | Experimental, Quasi-experimental, Observational |
| Location | Different country, Same country, Different country in the same region |
| Impact | $-5$, 0, $+5$, $+10$ percentage points |
| Organization | Government, NGO |
| Sample Size | 50, 3,000, 15,000 |

ten insignificant, even though there is evidence that this attribute is correlated with effect size (Bold et al., 2018; Cameron et al. 2019; Vivalt, 2020). Policy-makers at World Bank workshops exhibited a marked preference for studies with larger effects. Policy-makers cared about all attributes; researchers, however, notably did not put weight on any attributes except those related to sample size and methodology. Practitioners appear to fall somewhere between policy-makers and researchers for many of the attributes. They put particularly high weight on methods and sample size in the IDB sample, while still valuing the impact evaluation location.[6]

Appendix Table B4 presents results of an alternative specification using a mixed logit, with "impact" being considered fixed and all other variables assumed to be normally distributed. Again, results are comparable to the results from the conditional logit.

Overall, these results are consistent with the story in which policy-makers place more weight on factors relating to external validity than researchers.

[6]As before, we consider only the "pre" workshop IDB responses. The differences in the pre and post workshop responses are included in Appendix Table B5

Table 6: Seeking Research Results by Type of Respondent

| | *World Bank* | | | *IDB* | |
|---|---|---|---|---|---|
| | Policy-maker (1) | Practitioner (2) | Researcher (3) | Policy-maker (4) | Practitioner (5) |
| Impact | 1.053*** | 1.035* | 1.014 | 1.023* | 1.011 |
| | (0.017) | (0.018) | (0.021) | (0.012) | (0.019) |
| Quasi-Experimental | 1.625** | 2.180*** | 4.267*** | 1.331* | 1.527* |
| | (0.341) | (0.469) | (1.294) | (0.224) | (0.383) |
| Experimental | 2.473*** | 2.728*** | 8.869*** | 1.371** | 2.327*** |
| | (0.592) | (0.677) | (3.431) | (0.218) | (0.595) |
| Different country, same region | 1.563** | 1.492* | 1.077 | 1.556*** | 2.118*** |
| | (0.328) | (0.325) | (0.344) | (0.236) | (0.491) |
| Same country | 1.728** | 2.011*** | 1.386 | 2.363*** | 2.537*** |
| | (0.369) | (0.453) | (0.346) | (0.391) | (0.674) |
| Sample size: 3000 | 1.455* | 1.607** | 6.413*** | 2.007*** | 3.095*** |
| | (0.313) | (0.380) | (2.481) | (0.325) | (0.723) |
| Sample size: 15000 | 1.656** | 1.372 | 6.946*** | 1.974*** | 4.680*** |
| | (0.358) | (0.309) | (2.578) | (0.321) | (1.280) |
| Government | 1.338* | 1.015 | 0.951 | 0.948 | 1.434** |
| | (0.208) | (0.167) | (0.209) | (0.106) | (0.243) |
| Observations | 209 | 206 | 180 | 394 | 233 |

This table reports the results of conditional logit regressions on which impact evaluation was selected, using odds ratios. The omitted categories are "Observational", "Different region", "Sample size: 50", and "NGO". The number of observations represents the total number of choices made across individuals. The IDB results use only the pre-workshop sample. Standard errors are provided in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

# 5 Discussion

The results raise a natural question: could the differential weights given to evidence by researchers and policymakers lead to different conclusions from the evidence?

We cannot answer this question definitively, but we can use impact evaluation results collected in the course of meta-analysis to consider how researchers and policymakers would have weighed various papers and what conclusions they might have drawn. Here we use AidGrade's data set of 635 impact evaluation results from 20 different types of interventions (e.g., conditional cash transfers). The data are described in more detail elsewhere (Vivalt, 2020). Each result is from a study on a particular intervention and on a particular outcome - we will call these "intervention-outcome combinations" (e.g., the effect of conditional cash transfers on enrollment rates).

Now suppose we apply the policy-makers, practitioners and researchers' respective willingness-to-pay estimates to the studies in the sample. With these data, we can ask: which studies, within each intervention-outcome combination, would they have been willing to pay the most for? And what might they conclude from these "top studies"? The idea is that if policy-makers and researchers would be more interested in the results of studies that tend to find certain effects, this might lead them to be more likely to view those results, and in turn they might come away with biased beliefs about the effects of certain programs. The implicit model here is that there is some search cost increasing in the number of papers found. We assume the cost of finding each study's results is the same and that each policy-maker has a fixed budget to spend on obtaining new evidence. Then the ranked order of a policy-maker's preferred impact evaluations determines which they view. Recognizing that some policy-makers and researchers might be exposed to more literature than others, we do this analysis for different sets of "top" studies, namely, the top 5 studies within an intervention-outcome combination, the top 10 studies, and the top 20 studies. For the sake of having some study results that are not included in this top set, we restrict attention to those intervention-outcome combinations that contain results from at least 6, 11, or 21 papers, respectively, for which results can be standardized and we have full information on program implementer, sample size, and methods used.[7] This leaves us with a list of at most 21 intervention-outcome combinations across four

[7]Not all studies report full information; program implementer is particularly frequently unclear. For the sake of these analyses, we must restrict attention to those for which we have full information.

types of interventions, provided in the appendix.

For this back-of-the-envelope calculation, we leverage willingness-to-pay measures from the second experiment on seeking information from impact evaluations. The reason is three-fold. First, this experiment has a clearer conceptual link to the exercise: the premise is that policy-makers and others are seeking out information and may only view a restricted subset of results rather than putting some weight on all results. Second, in this experiment we have results for policy-makers, practitioners and researchers from the same setting, which makes direct comparisons of willingness-to-pay measures more compelling. Finally, it is only in the second experiment that we can connect data that relates to both internal and external validity.

The study characteristics that were considered in this experiment map closely to the study characteristics in AidGrade's data set. We have information on methods used, sample size and program implementer. We have impact in terms of effect sizes and in terms of raw units (e.g. enrollment rates in percentage points). For greater comparability across intervention-outcome combinations, we use effect sizes.[8] We also have information on where the study was done, but we cannot straightforwardly use it as different policy-makers would have been thinking of different countries as their "target" country in the discrete choice experiment, and we cannot estimate individual-specific willingness-to-pay measures with any precision. For this reason, we focus on the other factors, some of which reflect internal validity and some of which reflect external validity.[9] To generate the ranked list of studies within an intervention-outcome combination for each subgroup, we additively combine the estimated willingness-to-pay for each attribute. We have exactly the same categories coded for methods used and program implementer as in the experiment; since sample size is continuous, we interpolate willingness-to-pay estimates between the three sample sizes considered in the experiment.[10] Finally, to generate policy-makers', practitioners' or researchers'

[8]Interventions in the data seldom capture impacts in terms of percentage points.

[9]Sample size might reflect both internal and external validity. Its primary importance in an impact evaluation might be in determining a study's power. However, if there is specification searching or publication bias, smaller studies might systematically find higher effects than larger ones.

[10]For example, policy-makers were willing-to-pay 7.27 percentage points of impact for results from a study done on a sample size of 3,000 rather than a sample size of 50; they were also willing to pay an additional 9.76 minus 7.27 for results from a study done on a sample size of 15,000 rather than a study done on a sample size of 3,000. To generate their willingness-to-pay per additional data point between 50 and 3000, we use: 7.27/(3000-50); to generate their willingness-to-pay per additional data point above 3000, we use: (9.76-7.26)/(15000-3000). We generate their total willingness-to-pay for each study's actual sample size and add this to their willingness-to-pay for other attributes of

"best guess" based on their respective subsets of studies, we take the simple average of study results within that set. The average absolute differences between these "best guesses" and the meta-analysis means from random-effects meta-analyses using all the data from the same intervention-outcome combination are provided in Table 7.

Interestingly, despite their differences, policy-makers and researchers come to roughly comparable estimates. This may be due to the fact that while treatment effects significantly vary by sample size or program implementer, the differences tend to be small in magnitude overall in the underlying data set (Vivalt 2020). Further, AidGrade's data set does not contain observational data, and the differences in treatment effects between RCTs and quasi-experimental studies are small and insignificant in this data set.

Table 7: Difference Between Policy-Maker, Practitioner, and Researcher "Top Studies" and Meta-Analysis Results

| | Policy-Maker | Researcher | Difference | p-value | N |
|---|---|---|---|---|---|
| 5 studies | 0.05 | 0.06 | -0.00 | 0.22 | 21 |
| 10 studies | 0.04 | 0.04 | -0.00 | 0.73 | 17 |
| 20 studies | 0.05 | 0.03 | 0.02 | 0.36 | 7 |
| | Practitioner | Researcher | Difference | p-value | |
| 5 studies | 0.07 | 0.06 | 0.01 | 0.20 | 21 |
| 10 studies | 0.05 | 0.04 | 0.01 | 0.26 | 17 |
| 20 studies | 0.04 | 0.03 | 0.01 | 0.36 | 7 |

This table shows the average absolute difference between random-effects meta-analyses and the average effect sizes among studies preferred by policy-makers, practitioners and researchers according to the earlier willingness-to-pay estimates. Results are shown for the "top 5", "top 10", and "top 20" preferred studies within each intervention-outcome combination. Results are restricted to those intervention-outcome combinations that have results data from more studies than those used to make the prediction; "N" captures the number of intervention-outcomes under consideration for each row.

Considering the meta-analysis mean perhaps biases the exercise in favor of researchers: after all, if policy-makers place more weight on factors associated with external validity, perhaps they would do better at predicting the results in a given

that study, assuming each dimension is independent.

context. On the other hand, in their own contexts policy-makers may be more liable to be subject to behavioral biases such as confirmation bias.

This points to a second way in which policy-makers and researchers seek evidence differently in a manner that could drive differences in beliefs: policy-makers attach far more importance to treatment effects. Looking within the World Bank sample, policy-makers are close to 4% more likely than researchers to select an impact evaluation if it has reported a 1 percentage point higher impact. This makes intuitive sense as the impacts may be more important to policy-makers than to researchers. If policy-makers were to randomly draw a study from AidGrade's data set on the impacts of conditional cash transfers on enrollment rates with the bias that we observe, they would on average draw a study which estimated a treatment effect of 7.1 percentage points in the World Bank sample or 6.5 in the IDB sample, respectively 16.4% and 7.0% higher than the mean treatment effect of 6.1 percentage points.[11] Since this experiment asked policy-makers about programs they may or may not be personally interested in, we would expect the difference to be still larger when policy-makers are considering programs they may implement themselves. This may help to explain why policy-makers tend to have larger estimates of the effectiveness of various programs than researchers do (de Andrade et al., 2014; Hirschleifer et al., 2014; Casey et al., 2019; Vivalt and Coville, 2020).

# 6 Conclusion

This paper explores how policy-makers, practitioners and researchers seek information from impact evaluations and weigh evidence from different kinds of impact evaluations against recommendations from local experts. Overall, we find that policy-makers place more weight on factors associated with external validity while researchers place more weight on factors associated with internal validity. These preferences can have startling implications: according to our willingness-to-pay estimates, if a policy-maker were considering a program aimed at increasing enrollment rates they would be willing to accept a program that had been shown to have a 6.3 percentage point lower impact if the program came recommended by a local expert. They would also be willing to accept a program shown to have a 4.5 percentage point lower impact if it had been evaluated in their own country.

[11] The median treatment effect is even lower at 5.1 percentage points.

These choices may or may not be appropriate. They would make sense if, for example, results from another setting were uncorrelated or only slightly positively correlated to the effect the program would have in their setting and if policy-makers were risk-averse. However, the numbers are larger than we might expect given evidence on how much different study attributes are correlated with effect sizes. This suggests that understanding policy-maker preferences is critical when designing studies that aim to influence policy decisions.

# References


AidGrade (2016). AidGrade Impact Evaluation Data, Version 1.3.

Allcott, H. (2015). Site selection bias in program evaluation. *Quarterly Journal of Economics 130*(3), 1117–1165.

Banuri, S., S. Dercon, and V. Gauri (2015). The Biases of Development Professionals. In *World Development Report 2015: Mind, Society, and Behavior*, pp. 179–191. The World Bank.

Banuri, S., S. Dercon, and V. Gauri (2017). Biased Policy Professionals. *Working Paper series, University of East Anglia, Centre for Behavioural and Experimental Social Science (CBESS)*.

Bold, T., M. Kimenyi, G. Mwabu, A. Ng'ang'a, and J. Sandefur (2018). Experimental Evidence on Scaling up Education Reforms in Kenya. *Journal of Public Economics 168*, 1–20.

Cameron, L., S. Olivia, and M. Shah (2019). Scaling up sanitation: Evidence from an RCT in Indonesia. *Journal of Development Economics 138*, 1–16.

Casey, K., R. Glennerster, E. Miguel, and M. Voors (2019). Skill versus voice in local development. *Working Paper*.

Clark, J. and L. Friesen (2008). Overconfidence in forecasts of own performance: An experimental study. *The Economic Journal 119*(534), 229–251.

de Andrade, G. H., M. Bruhn, and D. McKenzie (2014). A helping hand or the long arm of the law? experimental evidence on what governments can do to formalize firms. *World Bank Economic Review 30*(1), 24–54.

DellaVigna, S. and D. Pope (2018a). What Motivates Effort? Evidence and Expert Forecasts. *The Review of Economic Studies 85*(2), 1029–1069.

DellaVigna, S. and D. Pope (2018b). Predicting Experimental Results: Who Knows What? *Journal of Political Economy 126*(6), 2410–2456.

DellaVigna, S. and D. Pope (2019). Stability of Experimental Results: Forecasts and Evidence. *NBER Working Paper #25858*.

Eil, D. and J. M. Rao (2011). The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself. *American Economic Journal: Microeconomics 3*(2), 114–138.

Hirshleifer, S., D. McKenzie, R. Almeida, and C. Ridao-Cano (2014). The impact of vocational training for the unemployed: Experimental evidence from turkey. *Economic Journal 126*(597), 2115–2146.

Hjort, J., D. Moreira, G. Rao, and J. F. Santini (2019). How Research Affects Policy: Experimental Evidence from 2,150 Brazilian Municipalities. *NBER Working Paper*.

Ioannidis, J. P. A., T. D. Stanley, and H. Doucouliagos (2017). The power of bias in economics research. *Economic Journal 127*(605), F236–265.

Krueger, A. O. (1993). *Political Economy of Policy Reform in Developing Countries (Ohlin Lectures)*. The MIT Press.

Kuzmanovic, B., A. Jefferson, and K. Vogeley (2014). Self-specific optimism bias in belief updating is associated with high trait optimism. *Journal of Behavioral Decision Making 28*(3), 281–293.

Langer, E. J. (1975). The Illusion of Control. *Journal of Personality and Social Psychology 32*(2), 311–328.

Liu, X., J. Stoutenborough, and A. Vedlitz (2016). Bureaucratic Expertise, Overconfidence, and Policy Choice. *Governance 30*(4), 705–725.

Maniadis, Z., F. Tufano, and J. A. List (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review 104*(1), 277–290.

Mobius, M., M. Niederle, P. Niehaus, and T. Rosenblat (2011). Managing self-confidence: Theory and experimental evidence. *Working Paper National Bureau of Economic Research*.

Moutsiana, C., N. Garrett, R. C. Clarke, R. B. Lotto, S.-J. Blakemore, and T. Sharot (2013). Human development of the ability to learn from bad news. *Proceedings of the National Academy of Sciences 110*(41), 16396–16401.

Nellis, G., T. Dunning, G. Grossman, M. Humphreys, S. D. Hyde, C. McIntosh, and C. Reardon (2019). *Information, Accountability, and Cumulative Learning*, Chapter Learning about Cumulative Learning: An Experiment with Policy Practitioners. Cambridge University Press.

Ortoleva, P. and E. Snowberg (2015). Overconfidence in Political Behavior. *American Economic Review 105*(2), 504–535.

Persson, T. and G. Tabellini (2002). *Political Economics: Explaining Economic Policy*. The MIT Press.

Rabin, M. and J. L. Schrag (1999). First Impressions Matter: A Model of Confirmatory Bias. *The Quarterly Journal of Economics 114*(1), 37–82.

Rogger, D. O. and R. Somani (2018). Hierarchy and Information. *Policy Research working paper, World Bank Group*.

Stuart, J. O. R., P. D. Windschitl, A. R. Smith, and A. M. Scherer (2015). Behaving Optimistically: How the (Un)Desirability of an Outcome Can Bias People's Preparations for It. *Journal of Behavioral Decision Making 30*(1), 54–69.

Vivalt, E. (2020). How much can we generalize from impact evaluations? *Journal of the European Economic Association 18*(6), 3045–3089.

Vivalt, E. and A. Coville (2020). Policy-makers consistently overestimate program impacts. *Working Paper*.

Vivalt, E. and A. Coville (2021). How do policy-makers update their beliefs? *Working Paper*.

Weinstein, N. D. (1987). Unrealistic Optimism about Susceptibility to Health Problems: Conclusions from a Community-wide Sample. *Journal of Behavioral Medicine 10*(5), 481–500.

Windschitl, P. D., A. M. Scherer, A. R. Smith, and J. P. Rose (2013). Why so confident? The Influence of Outcome Desirability on Selective Exposure and Likelihood Judgment. *Organizational Behavior and Human Decision Processes 120*(1), 73–86.

# Appendices

## A Forecasting Exercise Details

Given that it would be challenging to ask people to forecast the results of a conditional logistic regression (as both odds ratios and marginal effects may be challenging to guess), we simplified the exercise and asked individuals to merely forecast what share of the time they thought each group would choose the option with a given attribute, knowing that each option comprised several randomized attributes. These questions were of the form "If Program A were not recommended by a local expert, but Program B was, what percent of the time would a practitioner prefer Program B?" Since the levels of the attributes were randomized, an answer of 50% would imply that researchers thought policy-makers or practitioners were indifferent to whether the program was recommended by a local expert. Note that it is not the case that if policy-makers preferred studies recommended by a local expert they would always select Program B in this example, because other attributes were randomized at the same time and it is possible that Program A would have been preferred due to these other attributes. We noted at the top of each survey page that responses of 50% would indicate neutrality and answers of 100% were unlikely to be correct due to the randomization of other attributes. To aid researchers in providing forecasts that were appropriately scaled given this unknown source of noise, a reference value was given: the share of the time that practitioners chose the program that was associated with an estimated impact on enrollment rates of 10 percentage points over a program associated with an estimated impact of 0 percentage points (75%).[12] The forecasting survey text is provided in an online appendix.

We pre-specified that we would focus on forecasts from individuals with expertise in development (84 of the 159) and who spent at least two minutes completing the survey (84 out of 84), to screen out participants who simply clicked through without paying much attention to it. A handful of respondents appear to have still made

[12]The workshops had ended before forecasts were collected, but we had not yet incorporated information on participants from World Bank workshop rosters to fully classify them as policy-makers, practitioners and researchers and were using the self-reports to generate "preliminary" data. The 75% statistic turned out to be 74% after making adjustments based on these later data. As other researchers' responses were neither known nor provided, predictions of researcher responses were made *ex ante*. Again, no preliminary results for policy-makers or practitioners were publicly available beyond the single statistic provided.

inattentive forecasts, given that the question text for each section indicated that an estimate of 100% was implausible yet some individuals selected this value anyway (on average, 2.4 responses per question). We believe that answers below 50% are also likely to have been due to inattentiveness, as it seems unlikely for policy-makers as a whole to have actively preferred programs with impact evaluation results from other regions, and there were only 1.75 such responses on average per question. We thus exclude these estimates from the main analysis, along with those answers of 100%, however, we present the full distribution of forecasts for each question in Appendix Figure B2, and the mean forecast of each attribute does not change much if considering all responses.[13]

The full text of the forecasting survey is available at http://evavivalt.com/wp-content/uploads/Forecasting-Survey.pdf and will be archived at the Social Science Prediction Platform when the platform allows this feature.

[13]Results available upon request.

# B Additional Tables and Figures

Table B1: Weighing Research Results by Type of Respondent (Mixed Logit)

| | *Pooled* | | *World Bank* | | *IDB* | |
|---|---|---|---|---|---|---|
| | Policy-maker | Practitioner | Policy-maker | Practitioner | Policy-maker | Practitioner |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Mean | | | | | | |
| Impact | 1.075** | 1.131*** | 1.062* | 1.115** | 1.085 | 1.172*** |
| | (0.033) | (0.036) | (0.039) | (0.049) | (.) | (0.067) |
| Quasi-Experimental | 0.911 | 1.467 | 0.862 | 2.166* | 0.912 | 0.852 |
| | (0.239) | (0.438) | (0.238) | (0.869) | (0.521) | (0.258) |
| Experimental | 1.070 | 1.873* | 1.022 | 2.311 | 1.040 | 1.714 |
| | (0.249) | (0.659) | (0.220) | (1.188) | (0.539) | (1.053) |
| Different country, | 1.016 | 1.400 | 1.200 | 1.648 | 0.763 | 1.093 |
| same region | (0.212) | (0.368) | (0.317) | (0.668) | (.) | (.) |
| Same country | 1.355 | 2.055*** | 1.249 | 2.352** | 1.782 | 1.965* |
| | (0.316) | (0.491) | (0.352) | (0.806) | (.) | (0.686) |
| Recommended | 1.625*** | 1.181 | 1.409* | 1.085 | 1.940 | 1.384 |
| | (0.277) | (0.188) | (0.277) | (0.237) | (.) | (0.354) |
| Small CI | 1.250 | 1.555** | 1.159 | 1.619** | 1.336 | 1.557 |
| | (0.229) | (0.290) | (0.233) | (0.365) | (0.510) | (0.490) |
| SD | | | | | | |
| Quasi-Experimental | 1.592 | 1.812 | 1.455 | 2.108 | 1.767 | 0.987 |
| | (0.510) | (0.753) | (0.739) | (1.074) | (0.795) | (.) |
| Experimental | 1.321 | 2.950*** | 1.009 | 3.711*** | 1.500 | 2.713 |
| | (0.805) | (1.190) | (0.066) | (1.859) | (0.574) | (1.702) |
| Different country, same region | 1.225 | 1.980* | 0.939 | 2.974*** | 0.930 | 0.933 |
| | (1.082) | (0.717) | (0.777) | (1.195) | (.) | (.) |
| Same country | 1.937 | 1.091 | 1.768 | 0.971 | 1.712 | 1.806 |
| | (0.839) | (0.408) | (0.620) | (0.094) | (1.280) | (1.023) |
| Recommended | 0.907 | 1.008 | 1.020 | 1.025 | 1.484 | 1.022 |
| | (0.679) | (0.032) | (0.062) | (0.040) | (0.793) | (.) |
| Small CI | 1.645*** | 0.833 | 0.970 | 1.042 | 2.065 | 1.515 |
| | (0.252) | (0.381) | (0.121) | (0.074) | (.) | (0.748) |
| Observations | 239 | 267 | 143 | 156 | 96 | 111 |

This table reports the results of mixed logit regressions on which program was selected. Odds ratios are reported. The omitted categories are "Observational" and "Different region". The number of observations represents the total number of choices made across individuals. The IDB results use only the pre-workshop sample. Standard errors are provided in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table B2: Weighing the Significance of Research Results

| | *Pooled* | | *World Bank* | | *IDB* | |
|---|---|---|---|---|---|---|
| | Policy-maker | Practitioner | Policy-maker | Practitioner | Policymaker | Practitioner |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Impact | 1.027 | 1.032 | 1.042 | 1.047 | 1.004 | 1.021 |
| | (0.035) | (0.034) | (0.045) | (0.042) | (0.057) | (0.059) |
| Quasi-Experimental | 1.000 | 1.532** | 0.881 | 1.937** | 1.139 | 1.105 |
| | (0.208) | (0.319) | (0.234) | (0.520) | (0.401) | (0.384) |
| Experimental | 1.117 | 1.808*** | 1.043 | 1.971** | 1.260 | 1.690* |
| | (0.219) | (0.362) | (0.260) | (0.529) | (0.436) | (0.516) |
| Different country, same region | 1.046 | 1.431* | 1.201 | 1.502 | 0.880 | 1.337 |
| | (0.207) | (0.289) | (0.297) | (0.392) | (0.315) | (0.434) |
| Same country | 1.365 | 1.936*** | 1.242 | 2.048*** | 1.686 | 1.903* |
| | (0.268) | (0.378) | (0.314) | (0.518) | (0.561) | (0.644) |
| Recommended | 1.539*** | 1.236 | 1.354 | 1.091 | 1.862*** | 1.540 |
| | (0.227) | (0.177) | (0.259) | (0.198) | (0.444) | (0.404) |
| Small CI | 0.772 | 0.619 | 0.992 | 0.805 | 0.509 | 0.466 |
| | (0.254) | (0.206) | (0.420) | (0.353) | (0.292) | (0.253) |
| Significant | 2.044 | 3.811*** | 1.266 | 2.378 | 4.004 | 7.093** |
| | (0.973) | (1.782) | (0.776) | (1.387) | (3.425) | (6.022) |
| Observations | 239 | 267 | 143 | 156 | 96 | 111 |

This table reports the results of conditional logit regressions on which program was selected. Odds ratios are reported. The omitted categories are "Observational" and "Different region". "Significant" is a constructed variable; it was not explicitly provided in the options shown but could be worked out given the provided estimated impact and confidence interval. The number of observations represents the total number of choices made across individuals. The IDB results use only the pre-workshop sample. Standard errors are provided in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table B3: Weighing Evidence from Research Results vs. Local Experts, Pre/Post-workshop

| | *Both Rounds* | | *Either Round* | |
|---|---|---|---|---|
| | Pre | Post | Pre | Post |
| | (1) | (2) | (3) | (4) |
| Impact | 0.957 | 1.027 | 1.013 | 1.056 |
| | (0.070) | (0.060) | (0.039) | (0.041) |
| Quasi-Experimental | 1.514 | 1.453 | 1.112 | 2.057*** |
| | (0.660) | (0.595) | (0.270) | (0.522) |
| Experimental | 1.320 | 2.645** | 1.493* | 2.550*** |
| | (0.489) | (1.046) | (0.326) | (0.618) |
| Different country, same region | 1.587 | 4.529*** | 1.117 | 2.238*** |
| | (0.620) | (2.020) | (0.257) | (0.542) |
| Same country | 3.015*** | 4.149*** | 1.790** | 2.302*** |
| | (1.172) | (1.914) | (0.423) | (0.646) |
| Recommended | 1.845** | 1.692** | 1.691*** | 1.460** |
| | (0.560) | (0.441) | (0.295) | (0.262) |
| Small CI | 0.290* | 0.569 | 0.486* | 0.926 |
| | (0.185) | (0.350) | (0.187) | (0.354) |
| Significant | 19.207*** | 9.628*** | 5.490*** | 3.714** |
| | (19.791) | (8.207) | (3.166) | (2.086) |
| Observations | 87 | 96 | 207 | 216 |

This table reports the results of conditional logit regressions on which program was selected. Odds ratios are reported. These results focus on participants at the IDB workshops. "Both Rounds" refers to the participants who took both the "Pre" and "Post" survey; "Either Round" includes participants who only took one round. The omitted categories are "Observational" and "Different region". "Significant" is a constructed variable; it was not explicitly provided in the options shown but could be worked out given the provided estimated impact and confidence interval. The number of observations represents the total number of choices made across individuals. Standard errors are provided in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table B4: Seeking Research Results by Type of Respondent

| | *World Bank* | | | *IDB* | |
|---|---|---|---|---|---|
| | Policy-maker | Practitioner | Researcher | Policy-maker | Practitioner |
| | (1) | (2) | (3) | (4) | (5) |
| Mean | | | | | |
| Impact | 1.058** | 1.039 | 1.037 | 1.031** | 1.016 |
| | (0.026) | (0.029) | (0.032) | (0.016) | (0.024) |
| Quasi-Experimental | 1.659** | 2.358*** | 11.050*** | 1.426 | 1.545 |
| | (0.383) | (0.655) | (6.738) | (0.364) | (0.508) |
| Experimental | 2.594*** | 3.195*** | 34.380*** | 1.570 | 2.487** |
| | (0.756) | (1.321) | (33.770) | (0.582) | (0.887) |
| Same country | 1.714** | 2.129** | 1.723* | 3.827*** | 3.267*** |
| | (0.427) | (0.673) | (0.549) | (1.274) | (1.238) |
| Different country, | 1.570** | 1.500 | 1.243 | 2.020** | 2.490*** |
| same region | (0.301) | (0.390) | (0.799) | (0.645) | (0.692) |
| Sample size: 3000 | 1.472* | 1.664 | 18.253*** | 3.169*** | 3.590*** |
| | (0.294) | (0.660) | (15.358) | (1.135) | (1.268) |
| Sample size: 15000 | 1.757*** | 1.398 | 17.233*** | 2.998*** | 5.806*** |
| | (0.335) | (0.497) | (13.835) | (1.049) | (2.810) |
| Government | 1.420* | 0.973 | 0.984 | 1.067 | 1.515* |
| | (0.268) | (0.211) | (0.376) | (0.194) | (0.333) |
| SD | | | | | |
| Quasi-Experimental | 0.997 | 0.929 | 1.058 | 0.910 | 0.993 |
| | (0.005) | (1.091) | (0.234) | (0.224) | (0.047) |
| Experimental | 0.984 | 1.932* | 2.457 | 3.700*** | 1.341 |
| | (0.030) | (0.750) | (2.098) | (1.318) | (0.864) |
| Same country | 2.143** | 1.406 | 1.546 | 1.575 | 1.468 |
| | (0.684) | (0.584) | (1.230) | (0.763) | (0.668) |
| Different country, | 0.836 | 1.003 | 5.393*** | 2.789*** | 1.288 |
| same region | (0.485) | (0.013) | (3.268) | (0.838) | (0.603) |
| Sample size: 3000 | 0.996 | 1.005 | 3.244 | 2.115* | 0.994 |
| | (0.007) | (0.147) | (2.764) | (0.835) | (0.078) |
| Sample size: 15000 | 0.999 | 1.428 | 0.978 | 2.190** | 2.526** |
| | (0.005) | (0.693) | (0.072) | (0.752) | (1.165) |
| Government | 1.439 | 1.115 | 2.115 | 0.622** | 0.758 |
| | (0.406) | (1.586) | (1.245) | (0.138) | (0.239) |
| Observations | 209 | 206 | 180 | 394 | 233 |

This table reports the results of mixed logit regressions on which impact evaluation was selected. Odds ratios are reported. The omitted categories are "Observational", "Different region", "Sample size: 50", and "NGO". The number of observations represents the total number of choices made across individuals. The IDB results use only the pre-workshop sample. Standard errors are provided in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table B5: Seeking research results pre/post-workshop

| | *Both Rounds* | | *Either Round* | |
|---|---|---|---|---|
| | Pre | Post | Pre | Post |
| | (1) | (2) | (3) | (4) |
| Impact | 1.035** | 1.053*** | 1.024** | 1.038*** |
| | (0.014) | (0.014) | (0.010) | (0.011) |
| Quasi-Experimental | 1.297 | 1.568** | 1.467*** | 1.419** |
| | (0.223) | (0.290) | (0.195) | (0.209) |
| Experimental | 1.315 | 2.615*** | 1.740*** | 2.258*** |
| | (0.222) | (0.466) | (0.224) | (0.323) |
| Same country | 2.401*** | 2.894*** | 2.521*** | 2.681*** |
| | (0.435) | (0.567) | (0.343) | (0.414) |
| Different country, same region | 1.887*** | 1.554*** | 1.774*** | 1.484*** |
| | (0.299) | (0.256) | (0.217) | (0.194) |
| Sample size: 3000 | 2.044*** | 2.366*** | 2.293*** | 2.115*** |
| | (0.338) | (0.424) | (0.291) | (0.297) |
| Sample size: 15000 | 2.307*** | 3.143*** | 2.684*** | 2.535*** |
| | (0.391) | (0.570) | (0.360) | (0.364) |
| Government | 0.921 | 1.174 | 1.103 | 1.115 |
| | (0.110) | (0.152) | (0.099) | (0.113) |
| Observations | 373 | 396 | 664 | 558 |

This table reports the results of conditional logit regressions on which impact evaluation was selected. Odds ratios are reported. These results focus on participants at the IDB workshops. "Both Rounds" refers to the participants who took both the "Pre" and "Post" survey; "Either Round" includes participants who only took one round. The omitted categories are "Observational", "Different region", "Sample size: 50", and "NGO". "Significant" is a constructed variable; it was not explicitly provided in the options shown but could be worked out given the provided estimated impact and confidence interval. The number of observations represents the total number of choices made across individuals. Standard errors are provided in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table B6: List of Intervention-Outcome Combinations for Back-of-the-Envelope Calculation

| Intervention | Outcome | N |
|---|---|---|
| Conditional cash transfers | Attendance rate | 15 |
| Conditional cash transfers | Enrollment rate | 35 |
| Conditional cash transfers | Height-for-age | 7 |
| Conditional cash transfers | Labor force participation | 17 |
| Unconditional cash transfers | Enrollment rate | 13 |
| Deworming | Height | 16 |
| Deworming | Height-for-age | 14 |
| Deworming | Hemoglobin | 14 |
| Deworming | Mid-upper arm circumference | 7 |
| Deworming | Weight | 17 |
| Deworming | Weight-for-age | 12 |
| Deworming | Weight-for-height | 11 |
| Micronutrients | Birthweight | 7 |
| Micronutrients | Height | 28 |
| Micronutrients | Height-for-age | 33 |
| Micronutrients | Hemoglobin | 37 |
| Micronutrients | Mid-upper arm circumference | 17 |
| Micronutrients | Test scores | 9 |
| Micronutrients | Weight | 30 |
| Micronutrients | Weight-for-age | 31 |
| Micronutrients | Weight-for-height | 26 |

This table lists the intervention-outcome combinations with at least 6 studies on them in AidGrade's data set, after restricting attention to those studies whose results can be standardized and for which we have full information on the program implementer, sample size and methods used.

Figure B1: Example of a choice scenario

Now imagine that you need to provide a recommendation to a counterpart agency in your country on which of two programs to implement. A study was done on each program, with the results below. Please select which program you would recommend.

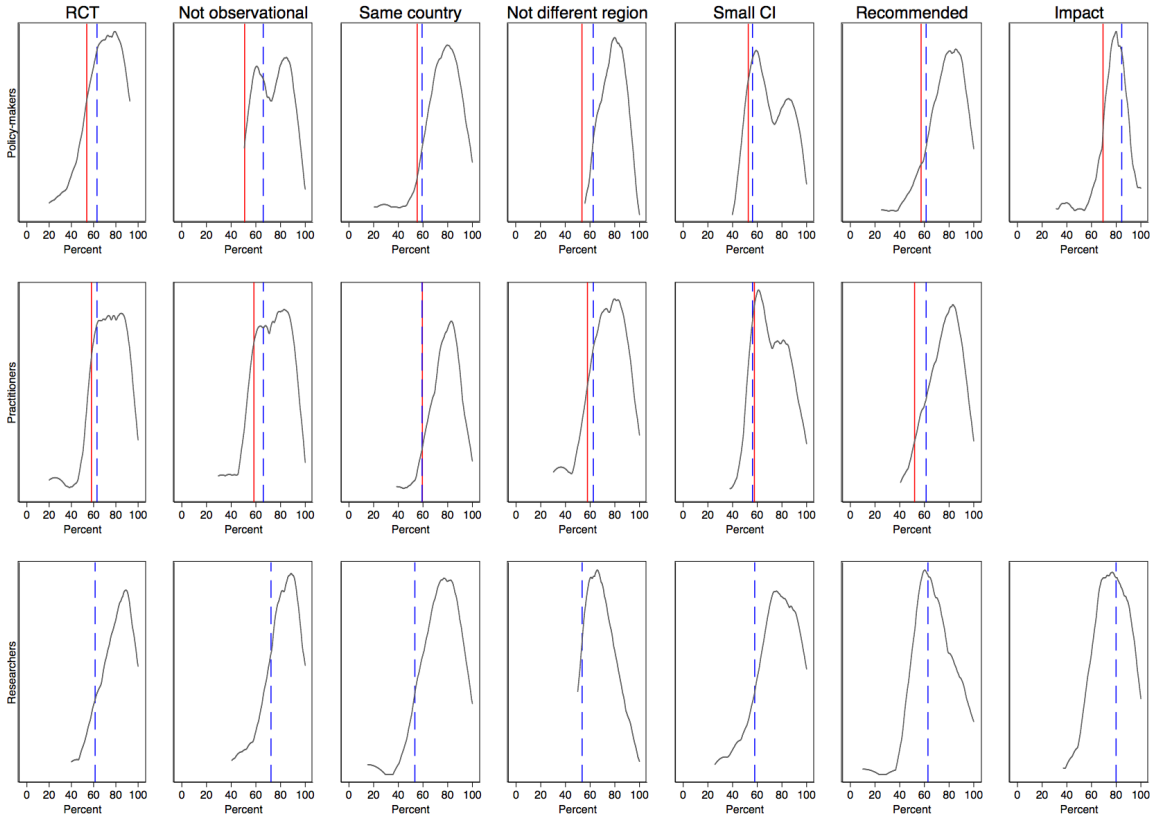
| | *Study on Program A* | *Study on Program B* |
|---|---|---|
| **Method** | Observational | Quasi-experimental |
| **Location** | A country in a different region | Same country |
| **Impact on enrollment rates, with margin of error (95% confidence interval)** | 0 percentage point, +/-10 percentage points | +10 percentage points, +/-1 percentage point |

A local expert tells you that they believe Program B would perform better in your context.

Which program do you recommend?

Program A ○ Program B ○

Figure B2: Distribution of Forecasts




This figure shows the distribution of forecasts made by researchers for policy-makers, practitioners, and researchers, respectively. The solid red line indicates the mean observed values for policy-makers and practitioners at the workshops. The dashed blue line indicates the mean observed values for those who took the online forecasting survey, based on the six question block each individual faced prior to making predictions. Some forecasts were obtained from policy-makers or practitioners both via targeted outreach to World Bank staff as well as through Twitter. 47 forecasters reported being practitioners or policy-makers, including 13 policy-makers; these are pooled due to the small sample size, according to our pre-analysis plan which specified analyzing policy-makers and practitioners separately if each contained at least 20 individuals.