Weighing the Evidence: Which Studies Count?

Eva Vivalt*

Aidan Coville†

Sampada KC‡

May 19, 2022

Abstract

We present results from two experiments run at World Bank and Inter-American Development Bank workshops on how policy-makers, policy practitioners and researchers weigh evidence and seek information from impact evaluations. We find that policy-makers and policy practitioners 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 were willing 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. Further, policy-makers and policy practitioners who had the most accurate forecasts of estimated program impacts were those who acted the most like researchers in seeking evidence, and vice versa.

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

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

‡University of British Columbia, samkc.student@ubc.ca. We thank Oscar Mitnik, Sebastian Martinez, and DIME for enabling the surveys to be run. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

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. One of the constraints in using evidence for policy-making is the availability of methodologically sound research on topics and in contexts that are relevant to the policy problem at hand. A surge of new studies in the last two decades has somewhat helped ease this constraint. At the same time, as the research base for a particular topic grows, opportunities for selective use of evidence increase. 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, this raises several questions: how do policy-makers weigh different kinds of research results or seek evidence to inform policy decisions? Do others, such as researchers, assess the same pool of evidence differently? If so, could these differences affect beliefs?

To explore these questions, we leverage a unique opportunity to run a set of experiments with policy-makers, policy practitioners and researchers invited to World Bank (WB) and Inter-American Development Bank (IDB) impact evaluation workshops. The WB workshops are designed as "matchmaking" events between government officials interested in impact evaluation and researchers who can help them develop an impact evaluation for one of their programs, while the IDB workshops are designed to more generally inform policy-makers about impact evaluation. Participants at these workshops include monitoring and evaluation specialists and program officers within government agencies who are in charge of certain programs and interested in impact evaluation; World Bank or IDB operational staff and a few other aid agency operational staff or program officers associated with those programs, such as technical advisors at USAID or the FCDO;[1] and researchers, both from academic institutions as well as some international organizations. Our focus on impact evaluation workshop participants is intentional: the attendees represent those directly involved in the evidence-policy cycle who are demonstrably interested in both the use and production of impact evaluations. These mid-level staff in government agencies (henceforth, *policy-makers*) and international organization staff and technical advisors (henceforth,

[1]USAID and the FCDO are the bilateral development aid agencies of the United States and United Kingdom, respectively.

policy practitioners) are often tasked with designing projects and advising higher-level policy-makers who will often not have time to engage with academic research directly. Our set of policy-makers also typically have the authority to make relevant decisions about particular programs, such as whether a program will be evaluated by impact evaluation and, if so, what methods it will use. Learning about these individuals' preferences could help us learn about the conditions under which different types of impact evaluations arise, which is important when thinking about external validity.

We run two experiments at these workshops. First, we leverage a discrete choice experiment to consider how policy-makers and policy practitioners *weigh* the value of different kinds of evidence they are presented with in selecting *programs*. These different kinds of evidence include 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, and our paper is the first to directly consider it. We find that, relative to researchers, who place a high weight on the methods of impact evaluations, policy-makers place a lower weight on impact evaluation methods and a higher weight on contextual factors, such as whether a local expert recommended the program. In other words, policy-makers value factors one might associate with external validity over factors one might associate with internal validity.

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.[2] Here, researchers participated in the experiment and forecast policy-makers', policy practitioners' and other researchers' responses. Researchers generally over-estimated how much weight policy-makers would place on each attribute but, after adjusting 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 researchers expect.

We also leverage a second set of discrete choice experiments to consider how policy-

[2]Forecasting survey sspp-2020-0002-v1, archived at: https://socialscienceprediction.org/s/b4ea2c.

makers, policy 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 found that a program had no effect in order to investigate the reasons for the lack of impact. However, all else equal, they would not want to choose a program with negligible effects when weighing which program to select. In this second experiment, apart from method and location, we consider the 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 policy 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 (de Andrade et al., 2014; Hirshleifer et al., 2014; Casey et al., 2019; Vivalt and Coville, 2020).

Consistent with this, we find suggestive evidence that policy-makers and policy practitioners who form more accurate beliefs about the effects of a program are those who seek evidence more like a researcher - and researchers who form more accurate beliefs seek information more like a policy-maker or policy practitioner. In one experiment, we asked policy-makers, policy practitioners and researchers to forecast the effects of different interventions alongside a discrete choice experiment, and the policy-makers and policy practitioners who made more accurate forecasts were the ones who put relatively less emphasis on factors associated with external validity, while the researchers who made more accurate forecasts were the ones who paid relatively less attention to factors associated with internal validity. This finding adds to the literature on forecasting that suggests individuals make more accurate forecasts when they do not have a strong commitment to one particular framework but are able to hold multiple factors in mind.[3] In our study, always choosing the impact evaluation from the most similar context or always choosing the one that used the most rigorous methods, regardless of the other attributes of the impact evaluations

[3]E.g., see work by Tetlock (2005), who builds on Berlin (1953) in classifying forecasters into two crude types: the hedgehogs, who "know one big thing," and the foxes, "who know many things." The foxes tend to perform better across many forecasting exercises.

in question, may indicate the kind of strong commitment to one framework that is associated with making worse forecasts.

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 relatively sparse. Banuri et al. (2015) demonstrate that policy practitioners may be subject to behavioral biases. Nellis et al. (2019) investigate how policy practitioners learn from meta-analysis results as opposed to individual studies. Mehmood et al. (2021) show that training policy-makers in basic econometrics can increase demand for higher quality research designs, while Toma and Bell (2022) show that decision aids can change the weight that policy-makers place on estimates of program impact. 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, we ask policy-makers to select not just *studies* but *programs*. As described earlier, an individual may select a study for its information value even when they would not want to implement the program that was evaluated in it, such as when a program had a negligible or negative effect. Second, 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*. Third, we 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. Program impacts provide a natural unit of analysis because they are analogous to a public budget: they are a real cost born by the public that depends on the choices of the policy-maker. This experimental design 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; further, 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 very large compared to the typical effects of popular programs that improve enroll-

ment rates.[4] 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 the accuracy of participant forecasts of impact evaluation results.

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, policy practitioners and researchers *weigh* evidence when selecting *programs*. We turn to the set of discrete choice experiments on how policy-makers, policy practitioners and researchers *seek* information from *studies*. 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, policy practitioners, and researchers. The workshops were each approximately one week long and were designed as "match-making" 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 involved in development programs; and academics and other researchers. We classify those from developing country governments as "policy-makers"; international organization operational staff and NGO or private sector employees as "policy practitioners"; and those in academia

[4]For example, in AidGrade's meta-analysis data, the median treatment effect of 36 conditional cash transfer programs on enrollment rates was 5.1 percentage points (AidGrade 2016).

Table 1: Response Rate at Workshops

Institution	Location	Year	Eligible Attendees	Surveyed	Response Rate
Experiment 1: Weighing Evidence on Programs					
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 from Studies					
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 two researchers' responses from the IDB workshop, 12 from the World Bank workshop in Athens and two from the World Bank workshop in Marrakesh. Due to this, researchers were 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 two known to be researchers and the true number of eligible participants is likely lower.

or those who either have peer-reviewed publications or else have "research" or "impact evaluation" in their job title as "researchers".

Policy-makers, policy practitioners and researchers were not restricted to attend 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 from studies and attendees at the workshops in Greece and Morocco were asked to participate in the discrete choice experiment on weighing information on programs.

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 policy practitioners for a week-long training on impact evaluation methods. Unlike the WB workshops, designing an impact evaluation in collaboration with researchers was not part of the agenda, so there were only two responses from researchers in total across these workshops. As a consequence, no results will be presented for researchers at the IDB for either experiment.

The workshop organizers emailed the survey link to the participants before the workshops began and emailed a link to a second (identical) survey at the end of the workshops. Our results focus on the responses obtained before the workshop began, as they may be closer to the typical preferences of policy-makers and policy practitioners. Participation was encouraged but voluntary. Attendees at the 2016 and 2017 workshops were asked to participate in the discrete choice experiment on seeking information from studies and attendees at the 2018 workshop were asked to participate in the discrete choice experiment on weighing information on programs.

3 Experiment 1: Weighing Evidence on Programs

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

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

were simply labeled as *Program A* and *Program B* and were intended to raise school 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.[5] A block consisted of six choice sets of two alternatives each. We randomized the blocks across respondents and the questions within blocks. Individual respondents were presented with one block each at the

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

World Bank workshops in Athens and Marrakesh and two blocks at the IDB. We also constructed 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.

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. Policy 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. Results for this experiment are limited to policy-makers and policy practitioners, as few researchers attended these particular workshops: there were only two researcher respondents in the IDB workshop, 12 at the WB workshop in Athens, and two at the WB workshop in Marrakesh. In total, these individuals made only 58 selections in the discrete choice experiment.

Results appear largely comparable across the World Bank and IDB pre-workshop samples.[6] 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.

Interestingly, 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 among policy practitioners (Appendix Table B1). This suggests that significance is being used as a heuristic.

[6]Table B2 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 policy practitioners, this makes intuitive sense: they may have been less familiar with impact evaluation methods *ex ante*.

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.

What do these results imply in terms of how policy-makers might make trade-offs between programs supported by different types of evidence? In Appendix Table B3, we present estimates of participants' willingness-to-pay in terms of estimated impact. In this table, we can see that policy-makers would be willing to accept a program with a 6.3 percentage point lower estimated impact if it came recommended 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 suggest 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.

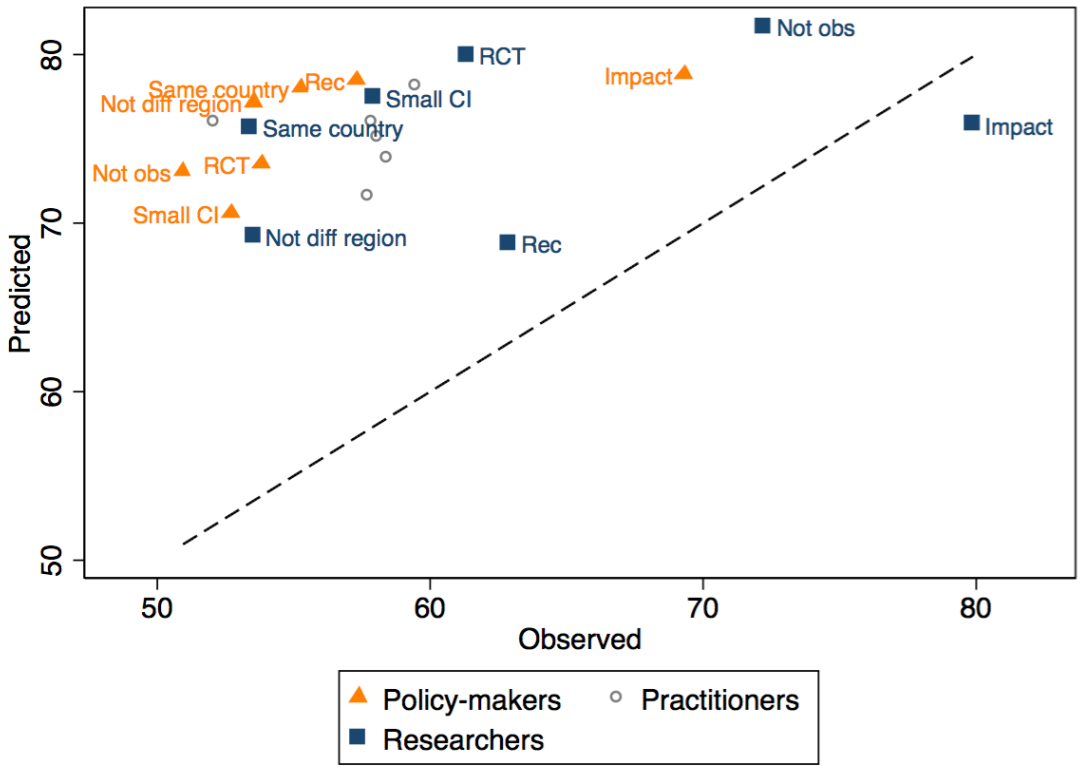
How do these results compare with researcher forecasts? We gathered forecasts from policy-makers, policy practitioners and researchers through the Social Science Prediction Platform between July 8, 2020 and August 17, 2020, as one of the first studies made available on the platform for others to forecast. 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, answering one randomly-selected block of six questions. This gave them familiarity with the questions the policy-makers and policy practitioners saw. After this exercise, they were asked to separately forecast the choices made by policy-makers and policy practitioners. Finally, participants predicted how researchers taking the forecasting survey weighed the different attributes when they completed the exercise. Details about the forecasting survey are provided in Appendix A, and the full forecasting survey can be viewed online.[7]

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 policy 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 re-

[7]Archived at: https://socialscienceprediction.org/s/b4ea2c.

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. "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.

sponses 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.

4 Experiment 2: Seeking Evidence from Studies

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 useful to them, rather than which of two *programs* they would prefer.[8] While weighing information on programs and seeking information from studies are clearly related, the attributes that are important to individuals when seeking information from studies could still differ from the attributes that are important when weighing which program to implement. For example, someone could wish to learn about an impact evaluation not because they think it did something

[8]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 4: 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

right but because they want to learn what not to do. Nonetheless, since exposure to different information could drive differences in beliefs and policy choice, it is 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 4.

More researchers attended the World Bank workshops at which this experiment was run and took the experiment, so in contrast to experiment 1, we can consider them as an additional subgroup. To highlight differences with researchers without generating a large number of interaction terms, we will consider both policy-makers and policy practitioners together as *policy professionals* and interact a dummy for being in this group with the other variables of interest. However, results for all subgroups are available in Appendix B.

4.2 Results

Table 5 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 as the target program, and from studies with large sample sizes. Interestingly, the coefficient on program implementer is not significant, even though

there is evidence that this attribute is correlated with effect size (Bold et al., 2018; Cameron et al., 2019; Vivalt, 2020). Policy professionals cared relatively more about studies being from the same country and relatively less about method used and sample size. Since no researchers attended the IDB workshops running this experiment, it must be recalled that the group of researchers comes solely from those participants at WB workshops and we cannot separately present results with these interactions using the IDB sample.[9] Results are further disaggregated in Appendix Table B5, and Appendix Table B6 provides a version of Table 5 that interacts the attributes with policy-maker status (grouping policy practitioners with researchers). In these robustness checks, policy-makers again put less weight on methods used and sample size.

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

4.3 Correlates of preferences towards impact evaluations

Differences in how individuals seek information could lead them to have different beliefs over time. In this section we consider some suggestive evidence that this is the case. In particular, the experiment run in Mexico included a module that asked respondents to forecast the impacts of various programs evaluated by impact evaluation.[10]

In this extra module, participants were first asked to indicate which intervention they were more familiar with: interventions involving SMS reminders; HIV awareness and prevention programs; or water and sanitation programs. Participants were then provided a list of titles, authors and dates of published impact evaluations and asked to select which papers they had heard of.[11] On average, participants reported having heard of 16.2% of the papers in the list. They were then provided with details about a randomly-selected impact evaluation on the topic with which they had expressed the most familiarity and were asked to estimate the program's effect.[12] If the impact eval-

[9] As before, we consider only the "pre" workshop IDB responses. The differences in the pre and post workshop responses are included in Appendix Table B4.

[10] This module was not included in the other workshops due to time constraints.

[11] The list consisted of: Leong et al., 2006; Jareethum et al., 2008; Liew et al., 2009; Mandirola, Guillen and Laguzzi, 2012; Odeny et al., 2012; Deb et al., 1986; Clasen et al., 2007; Kajubi et al., 2005; Patterson et al., 2008.

[12] Specifically, for every study, participants were provided with details about the study's location, sample, the intervention, the outcome variable, and the level of the outcome variable in the control

Table 5: Seeking Research Results by Type of Respondent

	Pooled (1)	World Bank (2)
Impact	1.026	1.014
	(0.018)	(0.021)
Quasi-Experimental	3.560***	4.267***
	(0.754)	(1.291)
Experimental	6.002***	8.869***
	(1.395)	(3.424)
Different country, same region	1.175	1.077
	(0.268)	(0.343)
Same country	1.446*	1.386
	(0.295)	(0.345)
Sample size: 3000	3.802***	6.413***
	(0.929)	(2.476)
Sample size: 15000	4.358***	6.946***
	(0.950)	(2.573)
Government implemented	0.886	0.951
	(0.153)	(0.209)
Policy professional * Impact	1.004	1.029
	(0.019)	(0.025)
Policy professional * Quasi-experimental	0.451***	0.442**
	(0.103)	(0.149)
Policy professional * Experimental	0.338***	0.290***
	(0.083)	(0.122)
Policy professional * Different country, same region	1.414	1.417
	(0.343)	(0.499)
Policy professional * Same country	1.476*	1.343
	(0.328)	(0.392)
Policy professional * Sample size: 3000	0.547**	0.243***
	(0.140)	(0.102)
Policy professional * Sample size: 15000	0.503***	0.220***
	(0.117)	(0.088)
Policy professional * Government implementer	1.293	1.220
	(0.243)	(0.300)
Observations	1243	595

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", as well as the equivalent categories interacted with policy professional status. "Policy professional" includes both policy-makers and policy practitioners. The number of observations represents the total number of choices made across individuals. The IDB results use only the pre-workshop sample and results for this sample cannot be presented separately as all IDB workshop participants are policy professionals. Standard errors are provided in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

Table 6: Seeking Research Results by Accuracy and Type of Respondent

	Pooled	Policy professional	Researcher
	(1)	(2)	(3)
Impact	1.020	1.015	1.062
	(0.028)	(0.033)	(0.083)
Quasi-Experimental	3.237***	3.733***	3.425
	(1.185)	(1.630)	(4.244)
Experimental	3.778***	4.094***	2.461
	(1.514)	(1.857)	(3.598)
Different country, same region	0.429**	0.494	0.256
	(0.181)	(0.234)	(0.395)
Same country	1.219	1.897*	0.239
	(0.371)	(0.676)	(0.278)
Sample size: 3000	2.711***	1.952*	1.333
	(0.966)	(0.778)	(1.461)
Sample size: 15000	2.752***	1.500	4.460
	(1.019)	(0.619)	(5.561)
Government implementer	0.842	0.672	6.791*
	(0.217)	(0.200)	(7.381)
Absolute error of forecast	1.004	1.006*	1.001
* Impact	(0.003)	(0.004)	(0.017)
Absolute error of forecast	0.971	0.924	1.273
* Quasi-experimental	(0.042)	(0.046)	(0.300)
Absolute error of forecast	1.014	0.973	1.532**
* Experimental	(0.045)	(0.046)	(0.270)
Absolute error of forecast	1.180***	1.185***	1.064
* Different country, same region	(0.066)	(0.074)	(0.201)
Absolute error of forecast	1.049*	1.045	1.123
* Same country	(0.030)	(0.033)	(0.175)
Absolute error of forecast	1.030	1.019	1.715**
* Sample size: 3000	(0.037)	(0.037)	(0.362)
Absolute error of forecast	1.020	1.039	1.536**
* Sample size: 15000	(0.043)	(0.049)	(0.305)
Absolute error of forecast	1.063**	1.095***	0.782
* Government implementer	(0.031)	(0.036)	(0.144)
Observations	252	168	84

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", as well as the equivalent categories interacted with the average absolute error of the forecasts made by the participant. "Policy professional" includes both policy-makers and policy practitioners. 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.

uation contained multiple treatment arms, they were asked to provide estimates for each treatment arm. While all the impact evaluations that they were asked about had already concluded, we restrict attention to the forecasts that participants provided for papers they claimed to have heard nothing about.[13]

All these studies had already been completed, but we do not anticipate that the participants would have been aware of their results. We calculate respondents' average accuracy as the average absolute difference between their forecasts and the estimated impacts.

Table 6 presents results. Interestingly, those policy professionals that put the most weight on location and implementer - factors associated with external validity - had the least accurate forecasts (larger absolute errors), while the researchers who put the most weight on sample size and method - factors associated with internal validity - had the least accurate forecasts. Given the small sample size, we do not wish to lean too hard on these results, but they are in line with the literature on expert forecasts and "cognitive styles", which finds that individuals who make decisions according to a single framework make less accurate forecasts than those who hold many frameworks in mind (Tetlock 2005; Mellers et al. 2015; Meynhardt et al. 2017). In our context, this trait might be associated with strict adherence to a rule of selecting studies according to some attribute. This interpretation remains speculative but suggests policy professionals and researchers can learn from each other in terms of placing weight on both internal and external validity.

5 Conclusion

This paper explores how policy-makers, policy 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

group at endline, and they were asked to predict the level of the outcome variable in the treatment group at endline. Appendix Figure B2 provides an example.

[13]It is theoretically possible that an individual may have heard of a program but not remember that they did, in which case their forecast would be biased towards being more accurate. Given that most individuals did not report having heard of many papers, and we expect that if anything participants may have exaggerated the number of papers they had heard about, we think this scenario unlikely. We cannot rule it out, but note that even in this case their accuracy would still be interesting to consider though it would have a different interpretation.

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. These trade-offs can be larger than the effect of the program. At the same time, policy-makers and researchers behaved more similarly than researchers expected in how they placed weights on different attributes of programs.

We also saw that the attributes that policy professionals and researchers pay attention to when seeking evidence from impact evaluations is correlated with their assessment about the effects of other programs. Policy professionals made less accurate forecasts if they put relatively more weight on factors associated with external validity than their peers, while researchers made less accurate forecasts if they put relatively more weight on factors associated with internal validity than other researchers. In other words, there is suggestive evidence that paying attention to both internal and external validity when seeking study results may lead to the most accurate beliefs, though further research is needed.

Given that individuals like those in our sample approve and implement many of the impact evaluations that form the evidence base on a number of topics in development economics, evidence on how they value study attributes can help us understand which impact evaluations may be selected to be run. This study also underscores the importance of understanding policy-maker preferences when designing studies that aim to inform 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)*.

Berlin, I. (1953). *The Hedgehog and the Fox: An Essay on Tolstoy's View of History*. Simon and Schuster.

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.

Clasen, T., T. F. Saeed, S. Boisson, P. Edmondson, and O. Shipin (2007). Household water treatment using sodium dichloroisocy anurate (nadcc) tablets: A randomized, controlled trial to assess microbiological effectiveness in bangladesh. *American Journal of Tropical Medicine and Hygiene 76*, 187–192.

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.

Deb, B. C., B. K. Sircar, P. G. Sengupta, S. P. De, S. K. Mondal, D. N. Gupta, N. C. Saha, S. Ghosh, U. Mitra, and S. C. Pal (1986). Studies on interventions to prevent eltor cholera transmission in urban slums. *Bulletin of the World Health Organization 64*(127).

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.

Gomez, P. and T. Meynhardt (2012). More foxes in the boardroom: Systems thinking in action. In S. N. Grösser and R. Zeier (Eds.), *Systemic Management for Intelligent Organizations*, pp. 83–98.

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.

Jareethum, R., V. Titapant, T. Chantra, V. Sommai, P. Chuenwattana, and C. Jirawan (2008). Satisfaction of healthy pregnant women receiving short message service via mobile phone for prenatal support: A randomized controlled trial. *Medical Journal of the Medical Association of Thailand 91*(458).

Kajubi, P., M. R. Kamya, S. Kamya, S. Chen, W. Mcfarland, , and N. Hearst (2005). Increasing condom use without reducing hiv risk: Results of a controlled community trial in uganda. *Journal of Acquired Immune Deficiency Syndromes 40*, 77–82.

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.

Leong, W. C., W. S. Chen, K. W. Leong, I. Mastura, O. Mimi, M. A. Sheikh, A. H. Zailinawati, C. J. Ng, K. L. Phua, and C. L. Teng (2006). The use of text messaging to improve attendance in primary care: A randomized controlled trial. *Family Practice 23*, 699–705.

Liew, S.-M., S. F. Tong, V. K. M. Lee, C. J. Ng, K. C. Leong, and C. L. Teng (2009). Text messaging reminders to reduce non-attendance in chronic disease follow-up: A clinical trial. *British Journal of General Practice 59*(569), 916–920.

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

Mandirola, H., S. Guillen, and P. Laguzzi (2012). It technologies to reduce the rate of missed appointments in the outpatients. *The 24th International Conference of the European Federation for Medical Informatics Quality of Life Through Quality of Information*.

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.

Mellers, B. A., E. Stone, P. Atanasov, N. Rohrbaugh, S. E. Metz, L. Ungar, M. M. Bishop, M. Horowitz, E. Merkle, and P. Tetlock (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied 21*(1), 1–14.

Meynhardt, T., C. Hermann, and S. Anderer (2017). Making sense of a most popular metaphor in management: Towards a hedgefox scale for cognitive styles. *Administrative Sciences 7*(33).

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.

Odeny, T. A., R. C. Bailey, E. A. Bukusi, J. M. Simoni, K. A. Tapia, K. Yuhas, K. K. Holmes, R. S. Mcclelland, and P. Braitstein (2012). Text messaging to improve attendance at post-operative clinic visits after adult male circumcision for hiv prevention: A randomized controlled trial. *Plos One 7*(9), E43832.

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

Patterson, T. L., B. Mausbach, R. Lozada, H. Staines-Orozco, S. J. Semple, M. Fraga-Vallejo, P. Orozovich, D. Abramovitz, A. de la Torre, H. Amaro, G. Martinez, C. Magis-Rodríguez, and S. A. Strathdee (2008). Efficacy of a brief behavioral intervention to promote condom use among female sex workers in tijuana and ciudad juarez, mexico. *American Journal of Public Health 98*(11), 2051–2057.

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.

Tetlock, P. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press.

Toma, M. and E. Bell (2022). Understanding and improving policymakers' sensitivity to program impact.

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*.

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 policy practitioner prefer Program B?" Since the levels of the attributes were randomized, an answer of 50% would imply that researchers thought policy-makers or policy 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 policy 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%).[14] The forecasting survey text is provided in an online appendix.

The forecasting survey was unincentivized. We pre-specified that we would focus on forecasts from individuals with expertise in development (84 of the 159) 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 re-

[14]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, policy 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*. No preliminary results for policy-makers or policy practitioners were publicly available beyond the single statistic provided.

spondents appear to have still made 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 B3, and the mean forecast of each attribute does not change much if considering all responses.[15]

The full forecasting survey is available at https://socialscienceprediction.org/s/b4ea2c.

[15]Results available upon request.

B Additional Tables and Figures

Table B1: 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 B2: 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 B3: 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.

Table B4: 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 implementer	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". 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 B5: Seeking Research Results by Type of Respondent

	World Bank			*IDB*	
	Policy-maker	Practitioner	Researcher	Policy-maker	Practitioner
	(1)	(2)	(3)	(4)	(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.

Table B6: Seeking Research Results by Type of Respondent

	Pooled	World Bank	IDB
	(1)	(2)	(3)
Impact	1.020**	1.031**	1.019
	(0.010)	(0.014)	(0.018)
Quasi-Experimental	2.249***	2.714***	1.818**
	(0.289)	(0.462)	(0.443)
Experimental	3.405***	3.871***	2.845***
	(0.484)	(0.750)	(0.719)
Different country, same region	1.559***	1.380*	2.121***
	(0.198)	(0.239)	(0.478)
Same country	1.854***	1.672***	2.762***
	(0.239)	(0.271)	(0.711)
Sample size: 3000	2.685***	2.627***	2.791***
	(0.362)	(0.491)	(0.648)
Sample size: 15000	2.901***	2.399***	4.450***
	(0.394)	(0.438)	(1.159)
Government implementer	1.139	1.027	1.402**
	(0.109)	(0.128)	(0.227)
Policy-maker * Impact	1.015	1.021	1.004
	(0.014)	(0.022)	(0.022)
Policy-maker * Quasi-experimental	0.653**	0.599*	0.732
	(0.116)	(0.162)	(0.217)
Policy-maker * Experimental	0.502***	0.639	0.482**
	(0.093)	(0.197)	(0.144)
Policy-maker * Different country,	1.016	1.132	0.734
same region	(0.174)	(0.308)	(0.199)
Policy-maker * Same country	1.161	1.034	0.856
	(0.207)	(0.277)	(0.262)
Policy-maker * Sample size: 3000	0.698**	0.554**	0.719
	(0.126)	(0.158)	(0.203)
Policy-maker * Sample size: 15000	0.677**	0.690	0.443***
	(0.122)	(0.195)	(0.136)
Policy-maker * Government implementer	0.937	1.302	0.676**
	(0.124)	(0.259)	(0.133)
Observations	1243	595	648

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", as well as the equivalent categories interacted with policy-maker status. 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.

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?

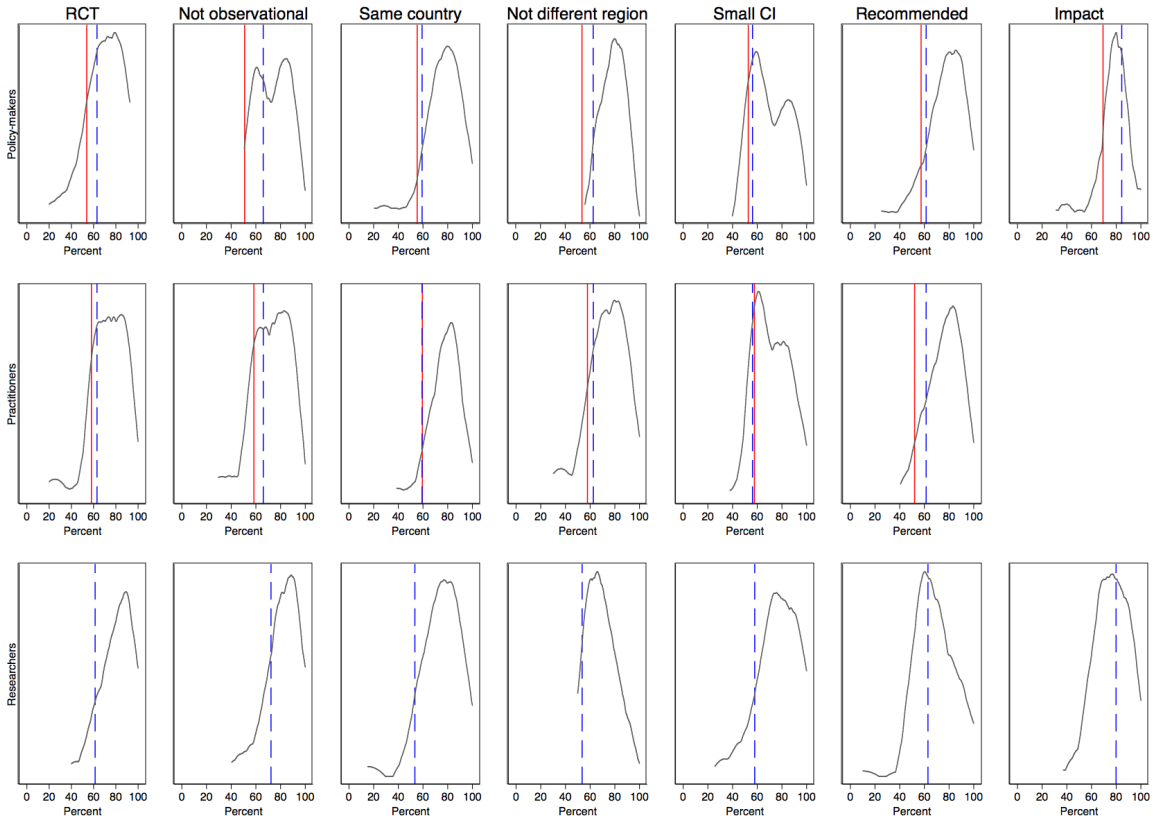
Program A ○ Program B ○

Figure B2: Sample Study Description

Location	Buenos Aires, Argentina
Sample	Men 18 years of age or older undergoing circumcision at clinics operated by the Ministry of Health. Men had undergone circumcision on the day of screening, owned a mobile phone, had the phone in their possession at the time of enrolment, and were able and willing to respond to a questionnaire administered by phone 42 days after circumcision. Men who reported prior or on-going participation in any other research study were ineligible.
Intervention	Intervention 1: Email reminder. Intervention 2: SMS reminder. Intervention 3: Phone call reminder. All reminders were sent out 72 hours before the scheduled appointment. SMS and email reminders were sent automatically while the phone call was delivered by operators trained to provide only the appointment details.
Outcome	Doctor appointment attendance
Control group level	71%

This figure shows one of the study descriptions that participants in the Mexico City workshop were shown before they were asked to forecast the results of an impact evaluation of the program. In this example, there are three treatment arms; individuals were asked to forecast all three, and the average error across treatment arms is calculated and used in running the analysis in Table 6.

Figure B3: Distribution of Forecasts



This figure shows the distribution of forecasts made by researchers for policy-makers, policy practitioners, and researchers, respectively. The solid red line indicates the mean observed values for policy-makers and policy 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 policy practitioners both via targeted outreach to World Bank staff as well as through Twitter. 47 forecasters reported being policy 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 policy practitioners separately if each contained at least 20 individuals.