Local Knowledge, Formal Evidence, and Policy Decisions

Eva Vivalt*

Aidan Coville†

Sampada KC‡

November 23, 2024

Abstract

How do policymakers value advice from local experts versus formal evidence from impact evaluations when making policy decisions? Using a discrete choice experiment conducted in collaboration with the World Bank and Inter-American Development Bank, we show that policymakers were willing to accept a program that had a 5.0 percentage point smaller estimated effect on enrollment rates if it were recommended by a local expert. They also preferred programs supported by evidence from a different region over programs supported by local evaluations only if the former had a 5.8 percentage point higher estimated impact. These premiums are large, surpassing the effects of many programs aimed at improving enrollment rates. This highlights the substantial weight that policymakers place on local evidence.

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

†Development Impact (DIME), World Bank, acoville@worldbank.org

‡University of British Columbia, samkc.student@ubc.ca. We thank Arianna Legovini for support and helpful feedback and Oscar Mitnik, Sebastian Martinez, and Silvia Velez Caroco for enabling the surveys to be run. Antonio Ortiz-Mena and Young Ju Nam provided excellent research assistance. Computational reproducibility for the results presented in this paper was verified by DIME Analytics. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. 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 significant increase in impact evaluations in recent years means that policymakers have more formal evidence than ever before to inform decisions on which policies to pursue. At the same time, policymakers often seem to rely on the recommendations of local experts and there is significant disagreement regarding how well local experts tend to forecast the effects of different programs (e.g., Bessone et al., 2021; Milkman et al., 2022; Iacovone et al., 2023). We consider how much policymakers weigh the recommendations of local experts relative to how much they weigh impact evaluation results and whether there are any features of impact evaluations that make them more likely to influence policymaker decisions, using a discrete choice experiment. We find that policymakers place significant weight on advice from local experts, but they also value evidence from impact evaluations in their country. This highlights the importance of generating evidence that is perceived as contextually relevant if impact evaluations are to inform policy decisions.

We conduct the discrete choice experiment at World Bank (WB) and Inter-American Development Bank (IDB) impact evaluation workshops. The WB workshops are meant to serve as "matchmaking" events that connect government officials, who are interested in using evidence from impact evaluation to inform their projects, with researchers supporting the design of such evaluations. The IDB workshops are intended to provide training to improve policymaker and policy practitioners' awareness and understanding of impact evaluations. Participants at both types of workshops are demonstrably interested in impact evaluation, while also being at the frontline of program and policy decisions, making them a particularly relevant target group for this study. We refer to this group of workshop participants, excepting researchers, as "policy professionals" and break this down further into government (*policymakers*) and international organization staff (*policy practitioners*). Individuals are identified and invited to the workshops based on the projects they are working on and their role within their respective organizations. The aim is to ensure the participation of teams with decision-making authority, program knowledge and day-to-day follow up capacity. Specifically, each project represented at the workshop includes five project team members on average: a high-level policymaker for decision power, a project lead from the government and an international organization project lead to ensure requisite knowledge about each project is available, and junior analysts or monitoring and

evaluation (M&E) officers to support follow-up activities. This team is paired with researchers for a week with the express objective of developing evaluations to inform project designs. This is a relevant sample because the typical WB/IDB-supported project is developed through the interactions of people in these types of roles.[1] A more detailed set of profiles for each type of policy professional and their role in project and policy decision-making can be found in the appendix.

We surveyed 190 policy professionals, providing them with descriptions of programs that are each associated with an impact evaluation result and asking them to choose between them. The impact evaluations attached to the programs differ by their identification strategy; location; impact; and the precision of the estimate. We compare how participants weigh these impact evaluation results relative to advice from a local expert.

We find that policymakers place a relatively high weight on contextual factors, such as whether a local expert recommended the program and whether the program was evaluated by an impact evaluation in their country. This may be due to concerns with how much rigorous evidence from one setting may be transportable to another (Pritchett and Sandefur, 2015; Vivalt, 2020), an important consideration when papers are still fairly geographically clustered (Leight, 2022). Policy practitioners also care about contextual factors but additionally pay attention to the internal validity of evidence from impact evaluations, preferring programs evaluated by RCTs and programs with results that were precisely estimated.

Our study builds on existing work exploring the decision-making processes of policymakers. Rogger and Somani (2023) find that bureaucrats in Ethiopia have beliefs about their constituencies that vary significantly from official statistics and that pro-

[1]Specifically, governments typically have priority targets or policy areas (such as increasing school enrollment) and will request donor support to finance activities that support achieving these outcomes. Initial project designs will typically resemble existing / past activities, and teams may draw on both tacit information and existing evidence from impact evaluation studies and other research (if it exists) on how effective these activities been and make adjustments based on an iterative process, incorporating different inputs that people/organizations preparing the project bring to the discussion. This culminates in a draft "project document" that outlines the actual components that will be implemented to achieve the higher-level outcomes, which is then reviewed by senior management and adjusted to develop a final project plan for sign off. Throughout this process, decisions about which components will form part of the project and the nature of these project components are influenced by information available to the project preparation team who need to decide how to use this information to adapt the project plan. The sample included in this study was purposely selected based on their role in this particular process that guides how the pipeline of WB/IDB-supported projects are developed.

viding evidence briefings can help reduce this gap. Nellis et al. (2019) consider how policy practitioners weigh meta-analysis results compared to results from individual studies. Banuri et al. (2019) highlights the prevalence of behavioral biases among policymakers while Toma and Bell (2023) examine the impact of decision aids on policy choice. Hjort et al. (2019) consider Brazilian mayors' willingness-to-pay for information from impact evaluations. Finally, in a companion paper, we consider which attributes policymakers, policy practitioners and researchers find attractive when searching for evidence (Vivalt et al., 2023).

In contrast to this past literature, we ask policy professionals to select *programs* rather than *studies*. Seeking out new evidence and weighing evidence when selecting a program are related but conceptually distinct. For example, policy professionals could seek information from an impact evaluation that found that a program had a negligible effect in order to investigate the reasons for the lack of impact. However, all else equal, they would not want to choose a program associated with negligible effects when weighing which program to select. While our experiment considers hypothetical choices, our focus on programs rather than studies is perhaps closer to the main item of interest, namely, how policy professionals make policy decisions. In making these decisions, policy professionals may still place some weight on study attributes, but they are not the only things they consider.

Our short paper is also the first to consider how policymakers weigh expert advice relative to impact evaluation results. This is relevant to the practical concerns of many working in development, as expert knowledge is a resource commonly used by policymakers (Morgan, 2014). More broadly, this paper relates to the concept of "tacit" vs. "explicit" knowledge, a distinction often made in the field of knowledge management. Tacit knowledge can be defined as knowledge held by an individual that is hard to formalize. It may be based on personal experiences and intuitions (Polanyi, 2009). In our experiment, impact evaluation results represent explicit knowledge - these results are formalized and can be communicated and understood by others. However, advice from a local expert could include elements of both tacit and explicit knowledge. For example, a local expert could draw on a number of formal sources of evidence in coming to their conclusions, and how they interpret past findings and integrate this information with their broader understanding and experience to form a recommendation involves their tacit knowledge. While tacit knowledge has been shown to be important in decision-making in a number of diverse fields (e.g.,

Podgórski, 2010; Hanna et al., 2014; Meisch et al., 2022), it is relatively understudied in economics, and this is the first paper to examine how policymakers weigh this type of evidence.

Finally, to quantify these trade offs, we consider how much participants would be willing to give up in terms of estimated impact in exchange for a program being supported by a certain kind of information. Program impacts provide a natural unit of analysis for assessing trade-offs because they are analogous to a public budget: they are a real cost born by the public that depends on the choices of the policymaker. Our experimental design allows us to say, for example, that policymakers would accept a conditional cash transfer program that was not recommended by a local expert, over one that was, only if such a program had at least a 5.0 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 5.8 percentage point higher estimated impact. These estimated impacts are very large compared to the typical effects of popular programs that improve enrollment rates.[2]

Our results highlight the importance of local knowledge for policy decisions and suggest that researchers looking to maximize their impact leverage appropriate settings and communicate their research findings to those local experts from whom policymakers may seek advice.

The rest of the paper proceeds as follows. First, we discuss the data used in the experiment. Then we describe and present results from the discrete choice experiment. Finally, we discuss the implications of our results.

2 Data

We surveyed 190 policy professionals at several World Bank and Inter-American Development Bank workshops, listed in Table 1 below. The context for this paper is that governments typically have priorities they wish to target and will request donor support to implement activities that support those outcomes. Teams working to

[2]For example, Vivalt (2020) finds that conditional cash transfers improve enrollment rates by 5.3 percentage points, and unconditional cash transfers by 3.5 percentage points. A more recent meta-analysis of unconditional cash transfer programs found an improvement of 4.1 percentage points (Crosta et al., 2024). This suggests that, in this experiment, policymakers would be willing to trade away much if not all of the estimated impact.

develop a project proposal may draw on a variety of sources of information, including expert advice and evidence from impact evaluations, if available. Throughout this process, decisions about project components are influenced by this information, but relatively little is known about how these factors are weighed by the different types of actors that engage in this policy process. The sample of workshop participants we focus on contains many of the types of actors that could plausibly use impact evaluations to inform policy decisions. Of the 190 policy professionals attending these workshops, we obtained responses from 156, representing a very high response rate of 82%. 81 policymakers and 75 policy practitioners participated in the survey. The following subsections describe each sample in more detail.

2.1 World Bank Sample

We surveyed attendees at WB workshops organized in Athens (September 2019), Marrakesh (December 2019), Bangkok (July 2023), and Dar es Salaam (August 2023). Workshop attendees consisted of policymakers and policy practitioners.[3] Each workshop was approximately one week long and was meant to facilitate connections between government staff and researchers. Policymakers and policy practitioners were matched with researchers and worked together over the course of the workshop to design a prospective impact evaluation that could be used for their program.

Workshops were attended by participants from around the world and we observed high response rates across World Bank workshops (91%). We expect this is in part because of the data collection approach within the workshop format: surveys were conducted as part of the program agenda. The designated time slot was early in the workshops to mitigate the possibility of experimenter demand effects.

[3]Participants are classified into policymakers and policy practitioners based on their survey responses. Specifically, participants self-identify as government employees, publicly elected or appointed officials, international development organization operational staff, NGO employees, private sector employees, researchers, or other. Government employees and publicly elected or appointed officials are classified as "policymakers". International development organization operational staff as well as a few NGO and private sector partners collaborating on projects are classified as "policy professionals". Too few researchers participated in these particular workshops to serve as a useful comparison group. 5 individuals who had selected "other" were classified into one of the substantive categories using the same decision rule as in Vivalt et al. (2023), and 6 others were re-classified based on an *ex post* review of the data that identified a mismatch between self-reported employment and actual employment. Our findings do not substantively change if we use the original self-classification (results available upon request).

Table 1: Response Rate at Workshops

Institution	Location	Year	Eligible Attendees	Surveyed	Response Rate
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
World Bank	Bangkok, Thailand	2023	30	27	0.90
World Bank	Dar es Salaam, Tanzania	2023	31	30	0.97
Total			190	156	0.82

The IDB rows include responses from the "pre" period and, in parentheses, the "post" period. The total is calculated using the total number of unique respondents across both rounds of the IDB survey (28). We exclude researchers from both the "eligible attendees" and "surveyed" counts, as too few researchers attended these workshops to be considered. Otherwise, all workshop participants are considered to have been eligible. Overall, response rates were very high.

2.2 IDB Sample

Participants were also recruited from a workshop organized in May 2018 at the IDB headquarters in Washington, DC. Like the WB workshops, this workshop ran for approximately one week, was attended by policymakers and policy practitioners, and focused on impact evaluation methods. This workshop, however, did not include matching participants with researchers to design an impact evaluation as the World Bank workshops did.

Participants were emailed the survey link by the workshop organizers before the start of the workshop and again after the workshop (for a second, identical survey). We focus on the survey responses collected before the start of the workshop, as these may be more representative of the typical preferences held by policymakers and policy practitioners.[4]

3 Method

A brief introduction informed participants that they would be asked questions that referred to conditional cash transfer programs. We also defined terms that some

[4]We do not have a sufficiently large sample to make before/after comparisons.

participants may not have been familiar with ("experimental", "quasi-experimental" and "observational" - see Figure A1).

Participants were then asked to repeatedly choose which of two programs they would prefer, being asked to make this decision for a randomly-selected set (or block) of six questions.[5] The programs were identified only as *Program A* or *Program B*, and each was associated with a study. These studies varied by method (experimental, quasi-experimental or observational); location (same country, different country in the same region, different country in a different region); the effect the study found (an increase in enrollment rates by 0, 5, or 10 percentage points); the precision of the estimate (a confidence interval of +/- 1 percentage point or +/- 10 percentage points); and whether a local expert recommended it. These attributes are summarized in Table 2.[6] Respondents saw one randomly-selected set (or block) of six questions each at the World Bank workshops in Athens, Marrakesh, Bangkok and Dar es Salaam, and two randomly-selected sets of six questions at the IDB. We used a fractional factorial design to select the levels of the attributes that were part of each choice option within each question and block to optimize power.

Our main results are analyzed using conditional logistic regression, clustering at the individual level, though we will also present results estimated using a mixed logit. Specifically, we estimate:

[5]E.g. "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." Appendix Figure A2 provides the introduction shown before the choice sets, and Figure A3 shows an example of a choice scenario participants might have faced. Participants at the IDB instead were asked to make this decision for two randomly-selected sets of six questions.

[6]We focus throughout on the effects of conditional cash transfers on children's enrollment rates in school as that is a common outcome of interest among many low and middle-income countries and there have been extensive studies on these programs in many settings, increasing the chance that policy practitioners would be familiar with them. 54% of our sample was from lower-income or lower-middle income countries where enrollment rates are a common target of development interventions. In addition, 34 percent of the sample was from Latin America, where net enrollment rates for secondary school continue to be a challenge in the region despite many countries being classified in the upper-middle-income category. As an illustration, the national average for the net secondary enrollment rate in 2018 was 91 per cent in Argentina and 78 percent in Colombia (World Bank, 2018).

Table 2: Attributes and Levels in the Discrete Choice Experiment

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

This table shows the different attributes used in the discrete choice experiment and their levels.

$$U_{ij} = \beta Impact_{ij} + \gamma Confidence\ Interval_{ij} + \delta Recommended_{ij} + \sum_{k=2}^{n} \zeta_k Method_{ijk} + \sum_{l=2}^{n'} \phi_l Location_{ijl} + \epsilon_{ij}$$

where U captures the utility of individual i selecting alternative j, $Impact$ is the estimated impact of the alternative, $Confidence\ Interval$ is a binary variable capturing whether the alternative had a small confidence interval, *Recommended* is a binary variable reflecting whether or not that alternative was recommended by a local expert, and indicators for the method of the study ("experimental" or "quasi-experimental") and location of the study ("same country" or "different country in the same region") are included in $Method$ and $Location$, respectively.[7]

The questions asked were hypothetical. We expect that this might tend to bias our results towards zero, since individuals would have less motivation to consider each question carefully. Since the study aims to elicit participants' unbiased beliefs and there are no clearly "correct" answers, we did not provide incentives that depended on participant responses. No incentives were provided for completing the survey, and the surveys were anonymous to reduce the potential for experimenter demand effects.

[7]Standard errors are clustered at the individual level as this was the level of the randomization.

4 Results

Table 3 presents results from the discrete choice experiment estimated using conditional logistic regression. Policymakers preferred programs with larger estimated treatment effects or programs that came recommended by a local expert, as well as those that had an impact evaluation from their country (Column 2). 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 (Column 3). They also preferred programs that came recommended by a local expert.[8]

Table 4 translates these results into estimates of how much policymakers and policy practitioners would be willing to pay, in terms of estimated impact, for programs with these different attributes. Policymakers would prefer a program recommended by a local expert even if it had an approximately 5.0 percentage point lower estimated impact on enrollment rates (Column 2), which is very large relative to the range of effects that programs typically have on enrollment rates.

Table A1 further breaks down responses across the World Bank and IDB pre-workshop samples and tests whether policymakers and policy practitioners put statistically significantly different weights on different attributes. Results appear largely comparable across these sub-samples, though policymakers put relatively less weight on large or precise estimates and do not emphasize programs with impact evaluations done in the same country as much as policy practitioners do.[9]

To further explore the importance placed on precise estimates, we construct an indicator 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. When we include this variable in the regressions, the

[8]Note that in order to avoid framing a program as *not* recommended, whether the program was recommended was displayed outside of the table that contained study attributes; this also effectively clustered the study attributes together and kept the non-study attribute separate, to make the distinction between study-related information and non-study-related information easier for participants to parse. While we cannot rule out that something about the format affects the weight policy professionals place on advice from a local expert, we do not expect it to drive the observed preference for local evidence overall given that policymakers also place so much emphasis on impact evaluation results coming from their country, and location was displayed within the table of impact evaluation characteristics.

[9]This could partially be explained by their generally putting less weight on impact evaluation results compared to the recommendation of a local expert, although this is speculative.

Table 3: Weighing Different Kinds of Evidence

	Pooled	Policymaker	Policy Practitioner
	(1)	(2)	(3)
Impact	1.07***	1.05***	1.10***
	(0.01)	(0.02)	(0.02)
Quasi-Experimental	1.11	0.97	1.31
	(0.13)	(0.15)	(0.22)
Experimental	1.40***	1.30*	1.55**
	(0.16)	(0.19)	(0.29)
Different country, same region	1.15	1.05	1.31*
	(0.12)	(0.15)	(0.20)
Same country	1.59***	1.33**	2.02***
	(0.15)	(0.17)	(0.26)
Recommended by local expert	1.24***	1.28**	1.21*
	(0.10)	(0.15)	(0.12)
Small C.I.	1.29***	1.06	1.61***
	(0.11)	(0.12)	(0.20)
Observations	818	407	411

This table reports the results of conditional logit regressions on which program was selected. Odds ratios are reported. "Impact" refers to the estimate of the effect associated with the program; "Quasi-Experimental" indicates whether the study associated with the program was quasi-experimental; "Experimental" indicates whether the study associated with the program was an RCT; "Different country, same region" indicates whether the study associated with the program was described as done in a different country in the same region; "Same country" indicates whether the study associated with the program was described as done in the same country; "Recommended by local expert" indicates whether the program was recommended by a local expert; "Small C.I." refers to the estimates having small confidence intervals. The omitted categories are "Observational", "Different region", and "Large C.I.". 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 4: Willingness to Pay (in Terms of Estimated Impact)

	Pooled	Policymaker	Policy Practitioner
	(1)	(2)	(3)
Quasi-Experimental	-1.55	0.66	-2.89
	(1.61)	(3.10)	(1.84)
Experimental	-4.83***	-5.26	-4.70**
	(1.84)	(3.42)	(2.18)
Different country, same region	-2.07	-1.03	-2.88*
	(1.57)	(2.95)	(1.74)
Same country	-6.70***	-5.77*	-7.55***
	(1.73)	(2.97)	(2.09)
Recommended by local expert	-3.11**	-4.99*	-2.01*
	(1.23)	(2.84)	(1.12)
Small C.I.	-3.67***	-1.21	-5.13***
	(1.24)	(2.21)	(1.49)
Observations	818	407	411

This table reports the results of conditional logit regressions in terms of the implied willingness-to-pay for programs with certain attributes. "Impact" is implicitly included in each estimate. For example, policymakers 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 5.0 percentage point higher estimated impact on enrollment rates. The number of observations represents the total number of choices made across individuals. Standard errors are provided in parentheses. Standard errors are provided in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

preference towards programs with studies having larger and more precisely-estimated effects mostly goes away, suggesting these preferences may be driven by significance (Table A2). However, a limitation of these results is that the study was not set up to explicitly test this hypothesis and the levels of the attributes that were used to construct this variable only very crudely capture significance.[10]

There are several potential explanations for policymakers' preferring programs supported by local experts. For example, policymakers may prefer programs recommended by a local expert because they worry about the transportability of research findings, particularly when evidence is not available from their country, and the local expert's knowledge could comprise an alternative form of local evidence; in other words, they could value the expert for their *local* knowledge. They could also appreciate the tacit knowledge of a local expert, who may have synthesized a number of data sources, potentially including but not limited to impact evaluation results.[11] Ultimately, we cannot distinguish between these competing hypotheses. However, given that policy professionals appear to care about two forms of local evidence - the recommendations from local experts and having impact evaluation evidence from the same setting - it is natural to wonder whether they view these as substitutes. The main results suggest this is not the case, since both variables are independently significant (Table 3). However, to further explore this issue, we interact whether a program was recommended by a local expert with whether impact evaluation results from the same country were displayed. Results in Table 5 suggest that policymakers do not put any more or less weight on advice from a local expert depending on whether an impact evaluation done in the same country is available, though it is possible that this is a function of limited power and the direction of the estimate for the interaction term is consistent with policymakers considering the two forms of local evidence to be partial substitutes.[12]

[10] In the knife-edge case of a point estimate of 10 and a confidence interval of 10, we consider that confidence interval to (barely) include 0 and hence be "insignificant".

[11] They could also plausibly trust specific experts, though we did not specify that the "local expert"was any particular individual in the experiment.

[12] Note that merely having impact evaluation results from the same country does not mean those impact evaluation results were promising. The average impact evaluation result was positive (representing an increase in enrollment rates of 5 percentage points), but the weight policymakers place on programs being recommended by local experts could in principle stem from those instances in which an impact evaluation showed no effect. The interaction term remains insignificant when all attribute levels are included, with the same caveat that the experiment was not designed to test for interactions (results available upon request).

Table 5: Having an Impact Evaluation from the Same Country is not Seen as a Substitute for Local Expertise

	Pooled	Policymaker	Policy Practitioner
	(1)	(2)	(3)
Impact	1.07***	1.05***	1.09***
	(0.01)	(0.02)	(0.02)
Same country	1.61**	1.51	1.78*
	(0.32)	(0.40)	(0.53)
Recommended by local expert	1.31**	1.44**	1.19
	(0.17)	(0.27)	(0.23)
Same country * Recommended by local expert	0.80	0.75	0.84
	(0.28)	(0.35)	(0.46)
Observations	818	407	411

This table reports the results of conditional logit regressions on which program was selected. Odds ratios are reported. The significance of the "Recommended by local expert" indicator and the insignificant interaction between "Same country" and "Recommended by local expert" in Column 2 suggests that policymakers do not put more or less weight on advice from a local expert depending on whether another form of local evidence - an impact evaluation done in the same country - is available. 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.

While our limited sample size prevents most analyses of heterogeneity, it may be informative to consider how much policymaker and policy practitioner preferences overlap. To this end, we use a mixed logit to estimate the same specifications as Table 3 but allowing effects to vary by individual (a "random effects" model). This model also inherently accounts for the fact that we have more observations for some participants than for others, as not all participants completed the full set of questions, and thus provides a further robustness check to the earlier regressions. Results are presented in Table A3. We observe some interesting patterns: while policy practitioners tend to put relatively more weight on factors relating to the internal validity of an estimate (method and the precision of the estimate) and policymakers put more weight on the recommendation of a local expert, there is substantial overlap in the distribution of their preferences (Table A3). For example, if we were to use the estimated standard deviation to construct the interquartile range (IQR) for the distribution of odds ratios across individuals, the IQR among policymakers for programs supported by experimental results would range from 1.02 - 1.83, while the IQR among policy practitioners would be 0.96 - 3.20. Similarly, the IQR among policymakers would be 0.78 - 1.59 for programs supported by studies with small confidence intervals, compared to 1.29 - 2.70 among policy practitioners. Looking at preferences towards programs recommended by a local expert, the IQR among policymakers is 0.91 - 2.28 and 1.23 - 1.42 among policy practitioners. While the typical policymaker cares more about this attribute than the typical policy practitioner, there is significant heterogeneity in preferences.

Research results serve as only one input into decision-making processes. While fully unpacking the role of different sources of information in policy decisions is beyond the scope of this short paper, we make a simple distinction between two commonly-used sources of evidence and consider how much weight policy professionals place on them. Because of our experiment's simplicity, there are necessary caveats. In particular, policy professionals in our setting choose between hypothetical programs and their decisions may differ under different circumstances. Further, the information we provide about these programs is necessarily limited in order to focus attention on the attributes of interest, and while the attribute levels were randomized, it remains possible that participants might have selected different options if other attributes had been shown. Nonetheless, we hope that this short paper highlights policymakers' strong demand for local evidence.

5 Conclusion

We find that policymakers place considerable weight on programs that are recommended by local experts as well as those with impact evaluations from their own country. The estimated policymaker preferences can have startling implications: according to our willingness-to-pay estimates, if a policymaker were considering a program aimed at increasing enrollment rates they would be willing to accept a program that had been shown to have a 5.0 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 5.8 percentage point lower impact if it had been evaluated in their own country. These trade-offs can be larger than the effect of many programs. These results suggest that unless research is seen by policymakers as valid in their target setting, policymakers are likely to choose alternative programs that may have lower estimated treatment effects but be from a more relevant-seeming context.

Given that individuals like those in our sample approve and implement many development programs, evidence on how they value the results that research provide can help us understand how to design studies such that they better feed into the evidence-to-policy pipeline. Our results imply that researchers aiming to improve policy should try to design studies to approximate the target context as closely as possible to maximize the chance that their results are taken up by policymakers.

References

Banuri, S., S. Dercon, and V. Gauri (2019). Biased policy professionals. *The World Bank Economic Review 33*(2), 310–327.

Bessone, P., G. Rao, F. Schilbach, H. Schofield, and M. Toma (2021). The Economic Consequences of Increasing Sleep Among the Urban Poor. *Quarterly Journal of Economics 136*(3), 1887–1941.

Crosta, T., D. Karlan, F. Ong, J. Rüschenpöhler, and C. R. Udry (2024). Unconditional cash transfers: A bayesian meta-analysis of randomized evaluations in low and middle income countries. *NBER Working Paper*.

Hanna, R., S. Mullainathan, and J. Schwartzstein (2014). Learning through noticing: Theory and evidence from a field experiment. *The Quarterly Journal of Economics 129*(3), 1311–1353.

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*.

Iacovone, L., D. McKenzie, and R. Meager (2023). Bayesian impact evaluation with informative priors: An application to a colombian management and export improvement program. *World Bank Policy Research Working Papers*.

Leight, J. (2022). Patterns of randomization, geographic concentration and funding in the recent development economics literature. *Working Paper*.

Meisch, S. P., S. Bremer, M. T. Young, and S. O. Funtowicz (2022). Extended peer communities: Appraising the contributions of tacit knowledges in climate change decision-making. *Futures 135*, 102868.

Milkman, K. L., L. Gandhi, M. S. Patel, H. N. Graci, D. M. Gromet, H. Ho, J. S. Kay, T. W. Lee, J. Rothschild, J. E. Bogard, and et al. (2022). A 680,000-person megastudy of nudges to encourage vaccination in pharmacies. *Proceedings of the National Academy of Sciences*.

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National academy of Sciences 111*(20), 7176–7184.

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.

Podgórski, D. (2010). The use of tacit knowledge in occupational safety and health management systems. *International Journal of Occupational Safety and Ergonomics 16*(3), 283–310.

Polanyi, M. (2009). The tacit dimension. In *Knowledge in organisations*, pp. 135–146. Routledge.

Pritchett, L. and J. Sandefur (2015). Learning from experiments when context matters. *American Economic Review 105*(5), 471–475.

Rogger, D. and R. Somani (2023). Hierarchy and information. *Journal of Public Economics 219*, 104823.

Toma, M. and E. Bell (2023). 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., A. Coville, and S. KC (2023). Seeking Evidence: Which Studies Do Policymakers Prefer? *Working Paper*.

World Bank (2018). World Development Indicators. https://databank.worldbank.org/source/world-development-indicators/Series/SE.SEC.NENR. Accessed: Oct. 10, 2024.

Appendices

A Additional Tables and Figures

Table A1: Comparing Policymakers and Policy Practitioners

	Pooled	World Bank	IDB
Impact	1.10***	1.09***	1.13***
	(0.02)	(0.03)	(0.05)
Quasi-Experimental	1.31	1.47*	0.91
	(0.22)	(0.29)	(0.25)
Experimental	1.55**	1.60**	1.46
	(0.29)	(0.35)	(0.61)
Different country, same region	1.31*	1.40*	1.06
	(0.20)	(0.28)	(0.22)
Same country	2.02***	2.14***	1.75***
	(0.26)	(0.34)	(0.36)
Recommended by local expert	1.21*	1.15	1.40
	(0.12)	(0.13)	(0.30)
Small C.I.	1.61***	1.64***	1.58*
	(0.20)	(0.24)	(0.39)
Policymaker * Impact	0.96*	0.96	0.95
	(0.03)	(0.03)	(0.06)
Policymaker * Quasi-Experimental	0.74	0.62*	1.08
	(0.17)	(0.16)	(0.58)
Policymaker * Experimental	0.84	0.83	0.72
	(0.20)	(0.23)	(0.43)
Policymaker * Different country, same region	0.80	0.83	0.69
	(0.17)	(0.21)	(0.26)
Policymaker * Same country	0.66**	0.60**	0.87
	(0.12)	(0.13)	(0.33)
Policymaker * Recommended by local expert	1.06	1.02	1.26
	(0.16)	(0.18)	(0.39)
Policymaker * Small C.I.	0.66**	0.61***	0.76
	(0.11)	(0.11)	(0.29)
Observations	818	611	207

This table reports the results of conditional logit regressions on which program was selected. Odds ratios are reported. Interaction terms are used to test whether policymakers weigh different attributes of studies differently. "Impact" refers to the estimate of the effect associated with the program; "Quasi-Experimental" indicates whether the study associated with the program was quasi-experimental; "Experimental" indicates whether the study associated with the program was an RCT; "Different country, same region" indicates whether the study associated with the program was described as done in a different country in the same region; "Same country" indicates whether the study associated with the program was described as done in the same country; "Recommended by local expert" indicates whether the program was recommended by a local expert; "Small C.I." refers to the estimates having small confidence intervals. The omitted categories are "Observational", "Different region", and "Large C.I.", as well as the equivalent categories interacted with the policymaker dummy. 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 A2: Significance is Important to Policymakers and Policy Practitioners

	Pooled	Policymaker	Policy Practitioner
	(1)	(2)	(3)
Impact	1.03	1.01	1.05*
	(0.02)	(0.02)	(0.03)
Quasi-Experimental	1.20	1.05	1.40*
	(0.14)	(0.16)	(0.24)
Experimental	1.45***	1.34**	1.61**
	(0.17)	(0.20)	(0.31)
Different country, same region	1.21*	1.11	1.37**
	(0.13)	(0.15)	(0.22)
Same country	1.61***	1.35**	2.04***
	(0.15)	(0.17)	(0.26)
Recommended by local expert	1.29***	1.32**	1.26**
	(0.10)	(0.16)	(0.13)
Small C.I.	0.73*	0.65*	0.88
	(0.14)	(0.16)	(0.24)
Significant	2.37***	2.15**	2.53**
	(0.64)	(0.83)	(0.97)
Observations	818	407	411

This table reports the results of conditional logit regressions on which program was selected. Odds ratios are reported. "Impact" refers to the estimate of the effect associated with the program; "Quasi-Experimental" indicates whether the study associated with the program was quasi-experimental; "Experimental" indicates whether the study associated with the program was an RCT; "Different country, same region" indicates whether the study associated with the program was described as done in a different country in the same region; "Same country" indicates whether the study associated with the program was described as done in the same country; "Recommended by local expert" indicates whether the program was recommended by a local expert; "Small C.I." refers to the estimates having small confidence intervals; "Significant" indicates whether a result might be perceived as significant according to its point estimate and confidence interval. The omitted categories are "Observational", "Different region", "Large C.I." and "Insignificant". 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 A3: Mixed Logit Model: Policymaker and Policy Practitioner Preferences Overlap

	Pooled		Policymakers		Policy Practitioners	
	Mean	SD	Mean	SD	Mean	SD
Impact	1.101***	0.131***	1.077***	0.114***	1.131***	0.150***
	(0.022)	(0.001)	(0.028)	(0.002)	(0.037)	(0.003)
	[1.01 - 1.20]		[1.00 - 1.16]		[1.02 - 1.25]	
Quasi-Experimental	1.107	0.060	0.955	0.328	1.350	0.310
	(0.154)	(0.058)	(0.181)	(0.070)	(0.296)	(0.105)
	[1.06 - 1.15]		[0.77 - 1.19]		[1.10 - 1.66]	
Experimental	1.526***	0.696***	1.370	0.432	1.751**	0.895*
	(0.239)	(0.062)	(0.264)	(0.145)	(0.457)	(0.219)
	[0.95 - 2.44]		[1.02 - 1.83]		[0.96 - 3.20]	
Different country, same region	1.156	0.387***	1.001	0.517*	1.401*	0.174
	(0.144)	(0.023)	(0.163)	(0.076)	(0.280)	(0.406)
	[0.89 - 1.50]		[0.71 - 1.42]		[1.25 - 1.58]	
Same country	1.832***	0.251	1.514**	0.296	2.430***	0.235
	(0.233)	(0.039)	(0.248)	(0.050)	(0.524)	(0.089)
	[1.55 - 2.17]		[1.24 - 1.85]		[2.07 - 2.85]	
Recommended by local expert	1.366***	0.386**	1.439**	0.682***	1.323**	0.109
	(0.140)	(0.035)	(0.261)	(0.055)	(0.166)	(0.084)
	[1.05 - 1.77]		[0.91 - 2.28]		[1.23 - 1.42]	
Small C.I.	1.380***	0.573***	1.113	0.530***	1.869***	0.548*
	(0.148)	(0.022)	(0.163)	(0.041)	(0.394)	(0.108)
	[0.94 - 2.03]		[0.78 - 1.59]		[1.29 - 2.70]	
Observations	818		407		411	

This table reports the results of mixed logit regressions on which program was selected. The "Mean" columns present the mean estimates, the standard error of the mean, calculated using the delta method (in parentheses), and the interquartile range (IQR) across individuals (in square brackets). The "SD" columns present the estimated standard deviations and the standard error of the estimates of the standard deviation (in parentheses). The "Mean" columns present results in terms of odds ratios for closer comparability with the rest of the paper, while the "SD" columns are presented in log-odds; the IQR is constructed from the mean and SD results and included to aid in interpreting the latter in terms of odds ratios. "Impact" refers to the estimate of the effect associated with the program; "Quasi-Experimental" indicates whether the study associated with the program was quasi-experimental; "Experimental" indicates whether the study associated with the program was an RCT; "Different country, same region" indicates whether the study associated with the program was described as done in a different country in the same region; "Same country" indicates whether the study associated with the program was described as done in the same country; "Recommended by local expert" indicates whether the program was recommended by a local expert; "Small C.I." refers to the estimates having small confidence intervals. The omitted categories are "Observational", "Different region", and "Large C.I.". The number of observations represents the total number of choices made across individuals. * p<0.10, ** p<0.05, *** p<0.01.

Figure A1: Description of Options

Before proceeding to view randomized choice sets and make selections, participants were shown the following definitions:

IMPORTANT DEFINITIONS FOR THE FOLLOWING QUESTIONS RELATING TO "METHOD":

"Method" refers to how treatment and control groups are determined in the study.

1. Experimental = a randomized control trial
2. Observational = a simple descriptive comparison of the outcomes of those who received an intervention and those who did not or a comparison of outcomes of those who received an intervention before and after the intervention
3. Quasi-experimental = a non-experimental approach that accounts for observable differences between control and treatment groups when comparing them

PLEASE READ THIS CAREFULLY BEFORE PROCEEDING.

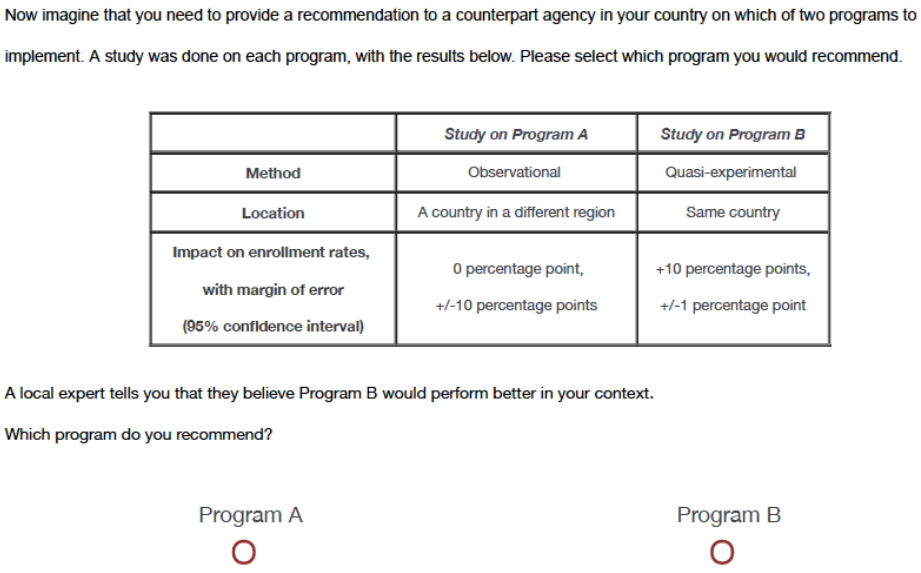
Figure A2: Introduction to Choice Scenarios

Participants were provided with the following introductory text immediately before being shown the choice scenarios:

Now imagine that you need to provide a recommendation to a counterpart agency in your country in charge of developing and implementing a new conditional cash transfer program. You will now be presented with different options to choose from.

Figure A3: Example of a Choice Scenario

This figure provides an example of a choice scenario participants might have faced.



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 ○

B Policy Professional Profiles

Here we present typical profiles of the five main types of policy professional present at the WB and IDB workshops. These types include: (i) High-level policymakers to provide decision authority and support links to national policy; (ii) Project leads from the government responsible for the overall implementation of the interventions/projects discussed; (iii) Junior analysts from the government responsible for data analytics and day-to-day project engagement between government and researchers; (iv) Project leads from the donor partner responsible for fiduciary oversight and quality control of the project; and (iv) Junior donor partner analyst responsible for day-to-day engagement between the project and research teams.

High-level government decision maker (policy engagement)
The high-level decision maker is a director-level civil servant or higher, heading a unit or department that typically includes multiple projects. They may have worked within the government for an average of about 15 years and are in charge of making decisions about a particular sub-sector. For example, in a Ministry of Water, this would be a director of rural water supply. As such they are well aware of broader policy around the sector and can speak to the context in which the particular operation is taking place, while also having influence on sector policy discussions. These participants held positions with anonymized job titles like: Director of Science, Technology and Environment, Chief Economist at National Development Bank, and Director at National Education Department.

Government project lead (operational engagement)
The project lead is ultimately responsible for the management of the specific project participating in the workshop. A typical World Bank-financed project will create a Project Implementation Unit (PIU) which will be housed within a ministry and be staffed by government officials that are responsible for implementing the project. The project lead is the point person for coordination with the Government and WB/Donor Partner, manages PIU staff, and bears responsibility for project delivery and key implementation decisions. These participants held positions with anonymized job titles like: Project Director at the National Agriculture Resource Board and Project Manager at the National Enterprises Agency.

Government junior analyst/monitoring & evaluation specialist (day-to-day engagement)

The junior analyst or monitoring & evalaution (M&E) specialist is hired into the PIU for the purpose of tracking the delivery of the project, managing different data sources (both administrative and primary data collected within the project), and reporting back on monitoring indicators used by the donor partner and government to track progress. They typically have some technical academic background, but do not have direct decision power on project or policy design. Instead, they provide inputs into the decision-making process. These participants held positions with anonymized job titles like: Officer at Ministry of Education and Vocation Training and Monitoring and Evaluation Specialist within the Ministry of Agriculture and Rural Development.

WB/Donor Partner project lead (project supervision)

The WB/Donor Partner sets up an internal project team responsible for the technical quality and oversight of the government-implemented project. The Task Team Leader (TTL) holds ultimately fiduciary responsibility of WB/Donor Partner financing and brings together and manages a team of technical experts (both in-house and as consultants) that provide ongoing supervision support to the project throughout its life cycle. A TTL usually rotates every 4 years, and thus is either involved in the design and early implementation, or close out of a project (since projects last 5 years on average). These participants held positions with anonymized job titles like: Education Specialist and Social Protection Specialist.

WB/Donor Partner analyst (day-to-day engagement)

The WB/Donor Partner team usually includes junior team members that provide more day-to-day technical support to the PIU in developing the logframe/monitoring targets, reviewing project progress and providing general support to the government as needed. These participants held positions with anonymized job titles like: Data Analyst, Research Analyst, and Operations Officer.