How Do Policymakers Update Their Beliefs?

Eva Vivalt*

University of Toronto

Aidan Coville

World Bank

March 12, 2023

Abstract — We present results from experiments run in collaboration with the World Bank and Inter-American Development Bank on how policymakers, policy practitioners, and researchers update their beliefs in response to results from academic studies. Initially, policymakers both believe development programs will have more positive results and are more certain about it than policy practitioners and researchers, despite reporting less familiarity with the programs. When participants are presented with the results of impact evaluations, we find evidence they update more on good news and are relatively insensitive to confidence intervals. We do not observe significant differences in biases between groups, and these biases cannot fully explain differences in beliefs.

*E-mail: eva.vivalt@utoronto.ca. We thank Sampada KC, Leo Kit Dai, Aguedo Solis Alonso, Marcos Pedreira Bernardo, Alma Bezares Calderon, Marinella Capriati, Timothy Catlett, Mark Engelbert, Jose Nicolas Rosas Garcia, Cesar Augusto Lopez, Huon Porteous, and Catalina Salas Santa for research assistance. We also thank Oscar Mitnik, Sebastian Martinez, and Silvia Velez Caroco, for enabling us to run the surveys. 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. AEARCTR-0001237.

1 Introduction

The last two decades have seen a large growth in the number of rigorous impact evaluation results that can be used to inform policy decisions. The implicit learning model that motivates the increase in impact evaluations is that policymakers have a set of beliefs that are informed by their prior experience, when policymakers are exposed to new evidence on the effectiveness of a particular intervention they update their beliefs, and this in turn may translate into changes in policy. To the extent that the new evidence provides a more accurate portrayal of the likely impacts of an intervention than a policymaker already has, this process has the potential to improve policy effectiveness. However, even if evidence exists and policymakers are interested in using it, they may not interpret the evidence correctly. In this case, new evidence would not help and, in some cases, could even hurt. Our paper explores this topic. In particular, we ask: (i) How do policymakers, policy practitioners and researchers' beliefs differ?; (ii) How do these groups update their beliefs when exposed to new research? Are they subject to particular behavioral biases?; and (iii) Can alternative ways of presenting research results overcome some of these biases?

To answer these questions, we leverage a unique opportunity to run a set of experiments on three groups of people involved in the generation and translation of research into policy decisions: policymakers, policy practitioners and researchers invited to World Bank (WB) and Inter-American Development Bank (IDB) impact evaluation workshops. The workshops are each approximately one week long and designed as "matchmaking" events between those involved in specific development programs and researchers to initiate future impact evaluations. Government officials are paired with researchers and tasked with designing a prospective impact evaluation for their program over the course of the week. Workshop participants include *policymakers* (program officers in government agencies, monitoring and evaluation specialists within government agencies and mid-level staff from line ministries), *policy practitioners* (World Bank or IDB operational staff, and other aid agency operational staff such as technical advisors at the US Agency for International Development (USAID) or the United Kingdom's Foreign, Commonwealth and Development Office (FCDO)), and a group of *researchers*, both from academic institutions and international organizations who participate in the workshop to support policymakers and policy practitioners in designing their prospective impact evaluations.

Our focus on participants attending these impact evaluation workshops is intentional and covers many of the key actors involved in generating and translating empirical research into policy decisions. How they interpret study results could influence different parts of the evidence-to-policy chain. Beyond producing evidence through impact evaluation, researchers are often tasked with translating and communicating that evidence for a policy

audience. Policy practitioners developing new programs are typically required to justify their proposed programs based on existing research. Finally, our sample of policymakers focuses on those who either have decision-making power over a particular program or who provide technical advice.[1] Workshop attendees are also demonstrably interested in impact evaluation and "evidence-based policy" more generally. We supplement the sample of workshop participants by running the experiment at the World Bank and the IDB headquarters in Washington, D.C.[2]

Our experiment is structured as follows. First, we present respondents with descriptions of interventions and elicit full prior belief distributions of the effects of these programs by asking participants to place probability weights on a range of possible treatment effects. We then show participants new evidence estimating the impacts of these programs and elicit their full posterior belief distributions. In the process, we randomly vary the point estimates and the confidence intervals attached to the results they observe.[3] This allows us to test for whether respondents update more on "good news" than "bad news" (asymmetric optimism) and whether they are insensitive to confidence intervals (variance neglect).[4]

While a number of other potential behavioral biases exist, we focus on these because of the implications they would have for policy decisions as the body of available evidence grows over time. Updating more on "good news" than "bad news" would result in decision-makers having inflated expectations about a policy's performance and a bias towards interventions with more evidence, regardless of whether the evidence is mixed. Given that some topics are covered by far more studies than others (Cameron et al., 2016), this could result in a persistent bias towards heavily-studied topics. If policymakers are relatively insensitive to confidence intervals, they will fail to differentiate between results with different levels of precision. This could compound the biases created by asymmetric optimism for two reasons. First, NGO-implemented pilot programs often have larger but more uncertain effects, and scale-ups often fare worse (Bold et al., 2018; Vivalt, 2020). This means that individuals who do not consider the precision of a result are likely to overestimate the effects of a program. Second, lower-powered studies, common in the empirical economics literature, are more likely to over-estimate the true underlying impact of the intervention, again with low precision (Gelman and Carlin, 2014; Ioannidis et al., 2017). This problem is particularly prevalent under publication bias (Brodeur et al., 2016). Overall, asymmetric optimism and

[1] While higher-level policymakers have more authority over policy decisions, they often do not have time to read about research findings themselves and may instead rely on briefings from individuals like those in our sample, making it a particularly relevant population to study.

[2] We also run the experiment on Amazon's Mechanical Turk (MTurk) for an additional comparison.

[3] There are sufficiently many studies on one of the interventions of interest that we can do this with real data in the first section of the experiment. In the second section of the experiment, participants are asked to consider hypothetical data to expand the set of results provided.

[4] Variance neglect is a novel bias closely related to extension neglect, in which individuals place less weight on the precision of results than a Bayesian would.

variance neglect have the potential to exacerbate gaps between beliefs and reality, and the gap may be widest when both these biases are present.

We have three main findings. First, we find that policymakers, policy practitioners and researchers have different prior beliefs. Policymakers expect larger effects and are more confident in their beliefs, even when they have less familiarity with the literature, while researchers have lower expectations but more uncertainty over them.

Second, we find evidence of both behavioral biases. The implications of these biases will depend on the initial beliefs and the evidence base, but in our setting these biases could lead respondents to expect a program to have almost a 10% greater effect than a Bayesian with the same prior beliefs would expect. Respondents are also more confident in their beliefs than a Bayesian would be: if they were to construct a 95% credible interval in which they thought the true effect would fall, that interval would be 31% narrower than that of a Bayesian. Interestingly, we do not find evidence of significant differences in biases between policymakers, policy practitioners and researchers, although these subgroups have significantly different prior beliefs. This suggests that the differences in prior beliefs that we observe may be more a function of what evidence individuals have previously been exposed to rather than differences in belief updating.

Finally, we show that how results are presented can influence how people update their beliefs. In two complementary experiments, we find that: (i) people are more responsive to information about precision (and thus closer to Bayesian) when presented with information on sample sizes rather than the equivalent confidence intervals; and (ii) providing more details about the distribution (i.e., including interquartile ranges together with confidence intervals) can increase responsiveness to the new information. Taken together, this suggests that the information that is presented is a potential tool that could mitigate some of the biases observed in this paper.

It can be challenging to find a setting to experimentally study policymaker decisions, and while there is an extensive literature on belief updating in other fields, evidence about policymakers is sparse. Hjort et al. (2021) 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, showing that policymakers value evidence on intervention effectiveness. However, Banuri et al. (2019) find that policy practitioners exhibit a number of behavioral biases in how they interpret information and make decisions. This suggests that even if policymakers both demand new evidence and change their decisions based on it, this is not guaranteed to translate into good policy. Approaches that have been tested to improve the belief updating process have included facilitating deliberation (Banuri et al., 2019), presenting meta-analysis versus individual results (Nellis et al., 2019)

and training policymakers on econometrics or providing decision aids (Mehmood et al., 2021; Toma and Bell, 2023). All of these approaches have had some success in improving updating. This suggests that new evidence, combined with interventions to improve how people interpret the evidence, may be able to improve policy effectiveness.

Our study advances this literature in several ways. First, our study is the first to collect the full distribution of priors and posteriors from policymakers.[5] This enables us to characterize the uncertainty in policymakers' beliefs, such as allowing us to observe that policymakers are more certain than researchers about their beliefs despite having less familiarity with the interventions. Second, it enables us to estimate not just whether policymakers are Bayesian, but how much they deviate from Bayesian updating. By parameterizing these differences we are able to explore the implications of these biases for belief updating, noting that the biases can be persistent even as the number of studies on a topic grows. Third, by experimentally varying the point estimates and confidence intervals of the data provided, we find that policymakers update more on "good news" than "bad news" and we are the first to document insensitivity to confidence intervals. Fourth, we are the first to compare the prior distributions and updating process across an important cross-section of actors involved in the evidence-based policy ecosystem - namely policymakers, policy practitioners and researchers.[6] Finally, we add to the literature on how updating can be improved by strategic choice of information, such as by presenting information on sample sizes. This has very practical implications for researchers wanting to ensure their results are taken up.

The rest of the paper proceeds as follows. First, we discuss the biases that we are studying. As others have done, we start from the premise that individuals are Bayesian updating and then introduce biases. In past literature, this has sometimes been referred to as "quasi-Bayesian" updating (Camerer et al., 2003). We then describe the sample and the experimental design. Finally, we present and discuss results, as well as the results from a small follow-up experiment.

2 Model

To more precisely state the biases of interest, we begin by describing a model of Bayesian updating, which we will then modify to introduce the behavioral biases of interest. Suppose a policymaker is deciding whether to implement a program. The program's effect if it were to be implemented in the policymaker's setting, θ_i, is unknown ex ante. The policymaker's

[5] Nellis et al. (2019) asked policy practitioners to provide a measure of uncertainty on a 3-point categorical scale, but this cannot be converted to a variance. Asking respondents to put weight in bins is considered a gold standard in terms of accurately estimating belief distributions (Delavande et al., 2011).

[6] Banuri et al. (2019), Nellis et al. (2019) and Toma and Bell (2023) more closely resemble our "policy practitioner" sample, according to their duties, while Hjort et al. (2021) and Mehmood et al. (2021) are more akin to our "policymaker" sample.

prior is that θ_i is normally distributed across settings, allowing for heterogeneous treatment effects:

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

where μ is the grand mean and τ^2 is the inter-study variance.

The policymaker has the opportunity to observe a signal about the effect of the program, Y_i with some normally distributed noise, *i.e.* $Y_i = \theta_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma_i^2)$. Y_i can be thought of as a point estimate in study i and has variance $\tau^2 + \sigma_i^2$, which we will write as v_i^2.

A person who is Bayesian updating will update their estimate of μ according to:

$$\mu_t = \mu_{t-1} + k(Y_i - \mu_{t-1}) \tag{2}$$

where $k = (v_{t-1}^2)/(v_{t-1}^2 + v_i^2)$. In other words, μ_t is a weighted combination of μ_{t-1} and the new information, Y_i, gained in period t. They will also update their estimate of the variance, so that:

$$v_t^2 = \frac{v_{t-1}^2 v_i^2}{v_{t-1}^2 + v_i^2} \tag{3}$$

Similar equations could be written for the case of no heterogeneity in treatment effects, in which case $\tau^2 = 0$ and $v_t^2 = \sigma_t^2$. This can be thought of as the appropriate model for when one is considering information from replications.

In our experiment, we focus on this latter case, explicitly framing the new information in the experiment as coming from replications. This is to avoid the estimation challenges posed by estimating $\tau^2 \neq 0$. In particular, if we were to introduce heterogeneous treatment effects, different people could build different mental models of how results depend on study characteristics and we would not be able to separately estimate the particular model they have in mind. To avoid this problem, even when we present information from different settings that might be subject to heterogeneous treatment effects we will not provide study details that could be used to build a more refined model; all studies are either described as replications or are otherwise "exchangeable".[7]

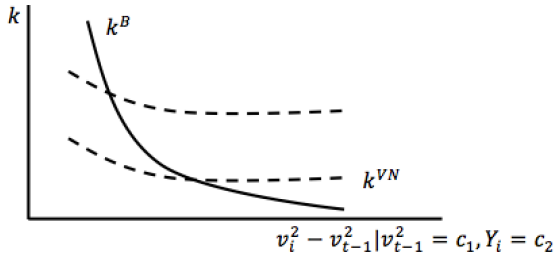
There are many ways in which individuals could deviate from Bayesian updating. We focus on two: asymmetric optimism and variance neglect.

2.1 Asymmetric Optimism

Our definition of asymmetric optimism follows the literature and means updating more on "good news" - information better than one's priors - than "bad news" (Eil and Rao, 2011). Several different kinds of asymmetric optimism with respect to good news may exist. One hypothesis is that respondents are optimistic with respect to point estimates,

[7]For more information on exchangeability, the interested reader is referred to Gelman et al. (2013).

Figure 1: Bayesian Updating vs. Variance Neglect



In this figure, while we do not know exactly where k^{VN} (dashed line) is in relation to k^B (solid line), we do know that its slope is less steep; in other words, for a given value of Y_i and v_{t-1}, different values of v_i result in values of k that are more similar to each other than the k of a Bayesian updater.

i.e. that they update more on point estimates higher than their prior mean than they do on point estimates symmetrically lower than their prior mean. For example, supposing μ_{t-1}=3, if we imagine that they alternatively receive the signal $Y_i = 1$ or $Y_i = 5$, we might expect them to update more when they receive the signal $Y_i = 5$.

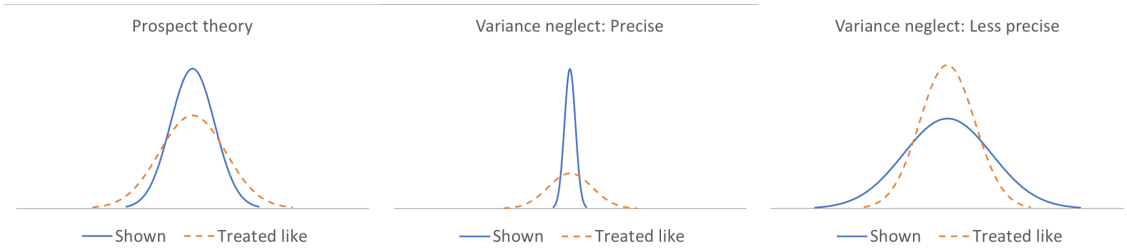
To model optimism formally, we adapt Rabin and Schrag's (1999) seminal paper on confirmation bias, which models confirmation bias as the misperception of a signal. In Rabin and Schrag's model, a person can observe one of two possible signals, but sometimes mistakes one signal for the other one. To expand this to a world in which many signals could be perceived, we might say that people observing a signal Y_i perceived that they saw $Y_i + \gamma$ for some $\gamma > 0$. This would result in a calculated k greater than the Bayesian k when presented with $Y_i > \mu_{t-1}$ and conversely a calculated k lower than the Bayesian k when presented with $Y_i < \mu_{t-1}$. If $\gamma = 0$, the policymaker would be unbiased.

2.2 Variance Neglect

For variance neglect, we only require that respondents pay less attention to the variance than they would if they were Bayesian updating. For example, consider the case in which respondents view data with alternatively small or large confidence intervals. If k_S^B (k_L^B) represents the k that a Bayesian updater would have if receiving a signal with small (large) confidence intervals, and k_S^{VN} (k_L^{VN}) represents the k that someone suffering from variance neglect would have upon receiving a signal with small (large) confidence intervals, $k_S^B - k_L^B > k_S^{VN} - k_L^{VN}$. Figure 1 illustrates.

Parameterizing this bias is straightforward: we say that individuals pay too little attention to either v_{t-1}^2 or v_i^2 when generating their estimate of μ_t, so that they update

Figure 2: Prospect Theory vs. Variance Neglect



This figure shows how data with varying precision might be treated under prospect theory versus under variance neglect. The left-most figure shows how someone might treat data under prospect theory, putting more weight in the tails (where there is small probability mass). The middle figure shows how someone might treat a precise estimate under variance neglect as more uncertain than it is. The right-most figure shows how someone might treat a less precise estimate under variance neglect as more precise than it is.

with:

$$k = \frac{v_{t-1}^2}{v_{t-1}^2 + (v_i^2 + \lambda)} \tag{4}$$

where λ reflects how much they underweight v_i^2 relative to v_{t-1}^2.

Variance neglect is closely related to, but distinct from, sample size neglect or, more broadly, extension neglect. In particular, sampling variance depends not just on the number of observations, but also on the standard deviation, and variance neglect could also apply to inter-study variation, though we do not test this latter possibility in this paper.

Prospect theory also bears similarities to variance neglect (Kahneman and Tversky, 1979). Under prospect theory, people overweight small probabilities and underweight large probabilities. They also treat gains and losses differently. The overweighting of small probabilities and underweighting of large probabilities should change how respondents treat normally distributed data. Given any normal distribution, respondents should act as though the distribution had larger variance; they should also act as though the distribution were skewed away from the side representing a loss. Prospect theory could result in variance neglect. However, there are other potential causes of variance neglect, and later we will see that prospect theory alone is inconsistent with some of our results. In particular, respondents do not always act as though the distribution had larger variance, and this will depend on the width of the confidence intervals they view, as in Figure 2.

Variance neglect is also related to the hot hand fallacy and the gambler's fallacy, nicely linked elsewhere (Rabin and Vayanos, 2010). Indeed, both the hot hand fallacy and the gambler's fallacy result in variance neglect. However, not all cases of variance neglect stem from the hot hand fallacy or gambler's fallacy. The situations in which hot hand fallacy

and gambler's fallacy classically arise are also descriptively different from the situation we are facing, in which policymakers do not view correlated data repeatedly but a few data points once.

The next sections provide more information on our data and methods.

3 Sample

We conduct this experiment with individuals attending World Bank or IDB workshops. Data were collected at six World Bank workshops run by the Development Impact Evaluation (DIME) research group. The workshops were conducted in Mexico City (March 2017), Lagos (May 2017), Washington, DC (May 2017, June 2017), Lisbon (July 2017), and Dakar (January 2019). The experiment was also conducted at two IDB workshops in Washington, DC in June, 2017, and May, 2018. Two other World Bank workshops (Mexico City, 2016, and Nairobi, 2016) were used as pilots to refine the survey questions and the prior/posterior elicitation mechanisms.

These conferences attract participants from around the globe. To accommodate more participants, the survey was translated into Spanish, and respondents had to be fully proficient in English or Spanish in order to participate. For the workshop in Dakar, French was also used.

Individuals were surveyed by enumerators during breaks in the workshops. Of 526 eligible attendees at the non-pilot workshops, 162 (31%) completed the survey. The main constraint was that the surveys could only be run during the typically twice-daily breaks in the workshops and during the lunch period.[8] Thus, this response rate represents essentially the maximum number of responses that could be gathered in the allotted time.[9] We may expect that those who managed to take the survey may have been particularly interested in taking it or quick to approach the enumerators during a break, but we have no reason to believe that this represents a substantially different population than the universe of conference attendees. Response rates by workshop are detailed in Table 1.

In addition to gathering data at these workshops, past workshop participants were contacted by e-mail and asked to participate via video conference. The response rate was much lower in the group contacted by e-mail; of 479 eligible past workshop attendees, 46 (10%) participated in the survey. Finally, participants were recruited at the World Bank's

[8] During the pilots, individuals were allowed to take the survey by themselves on tablets we provided. However, we changed approaches after the pilot in favor of one-on-one enumeration to reduce noise due to participants' lack of familiarity with operating the tablets and to increase attentiveness. After making this change, we still had overwhelming interest in the survey among attendees but, being limited to the breaks in the workshops, only managed to survey an average of 23 participants per workshop.

[9] Breaks were roughly the duration of the survey, and lunch might span 2-3 times the length of the typical break, depending on workshop timing.

headquarters and at the IDB's headquarters in Washington, D.C. A table was set up by the cafeteria and passers-by were able to take the survey with a trained enumerator. 125 responses were collected at the World Bank and 27 at the IDB over 24 days or 12 lunches, respectively;[10] enumerators covered lunch at the IDB but full or half-days at the World Bank. Summary statistics about the various recruitment strategies and the breakdown of participants by category (policymaker, policy practitioner, researcher) are provided in Table 2.

The experiment was set up to elicit two sets of priors and posteriors, so that with 400 respondents we would expect 800 priors and 800 posteriors. However, some respondents did not complete the entire survey, so we only present results for those who provided both a prior and a posterior for a given set of questions, resulting in 753 priors and posteriors from 378 respondents, or 94% of the total possible responses.

Table 1: Participants at Workshops

	Eligible Attendees	At Workshop	Post-Workshop	Response Rate
Mexico, March 2017	86	36	8	0.51
Nigeria, May 2017	74	36	1	0.50
Washington, DC, May 2017	47	13	2	0.32
Washington, DC, June 2017 (IDB)	62	10	0	0.16
Washington, DC, June 2017 (WB)	67	21	1	0.33
Portugal, July 2017	111	30	8	0.34
Washington, DC, May 2018 (IDB)	51	14	0	0.27
Senegal, January 2019	54	20	0	0.37
Total	552	180	22	0.37
Total (excluding IDB)	439	156	20	0.40

This table shows the number of people surveyed at each workshop and the total number of eligible attendees. Both values restrict attention to those who could be classified as "policymakers", "policy practitioners" or "researchers". In addition, to be eligible to take the survey, one had to have not taken it at a previous workshop (this was primarily a concern for DIME staff) and one had to speak a survey language fluently (English or Spanish or, for Senegal, English, French or Spanish). Two of the workshops were held by the IDB; all other workshops were held by the World Bank.

Finally, a set of responses was elicited on MTurk to explore whether the biases observed in our study sample represent a more general phenomenon. Details are provided in the Appendix.

[10] Excluding three responses from support staff at the World Bank and two responses from support staff at the IDB. These participants (IT staff, secretaries, lawyers) did not meet our inclusion criteria but we could not bar them from participating.

Table 2: Respondents by Recruitment Strategy

	Policymakers	Policy Practitioners	Researchers	Total
Workshops	0.38	0.31	0.32	180
Videoconference	0.16	0.29	0.54	68
Headquarters surveys	0.04	0.56	0.40	152
Total	0.21	0.40	0.39	400

This table shows the percent of respondents who could be classified as policymakers, policy practitioners and researchers by each recruitment strategy.

4 Experimental Design

4.1 Structure of the Experiment

At the beginning of the survey, participants were asked to answer some basic questions about their experience and familiarity with different programs. They were shown a video describing how to use the sliders to assign probability weights to different outcomes and were walked through a simple example about predicting the weather in order to be sure that they understood the exercise. At the end of this section, participants were asked if they understood and were only allowed to participate further if they stated that they did (Figure C1). Only one participant stated that they did not understand the instructions and was prohibited from continuing. Respondents continued on to answer questions in which they were shown a random selection of real-life data and asked to make a real-life allocation decision. They then were asked questions using hypothetical data.

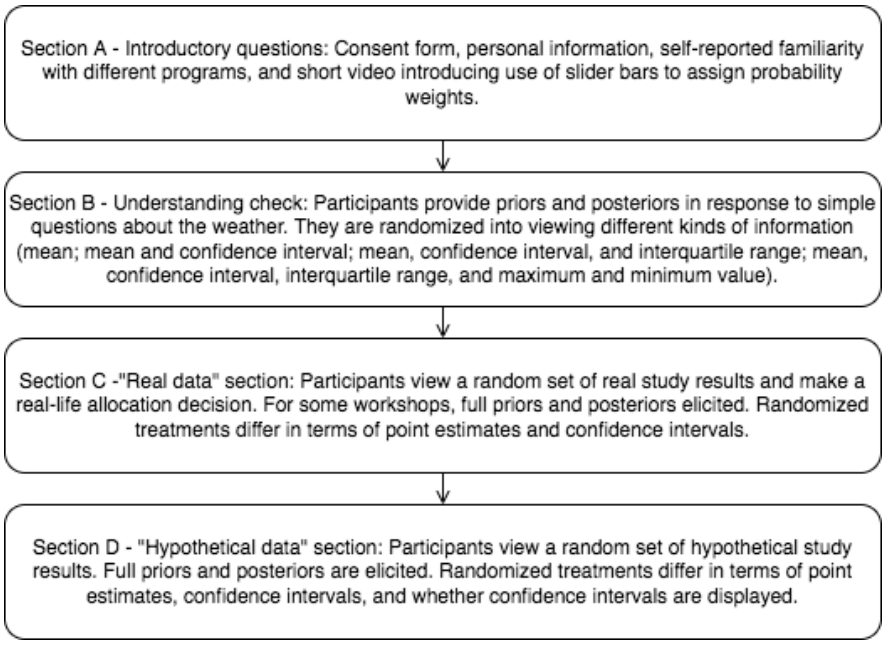
Figure 3 demonstrates the overall structure of the survey. Our results on asymmetric optimism and variance neglect will draw on the results from Section D; our results on how evidence from impact evaluations affects decision-making will draw from Section C; and our results on how individuals update their beliefs in response to different types of information will draw on the data generated by Section B.[11]

4.2 Real-Life Data and Decisions

Respondents were told that they would be shown real data from impact evaluations and that they would help decide how a small amount of external funds would be allocated - in

[11]It is in principle possible for the different portions of the experiment to be used to provide evidence on questions targeted by other sections. For example, at present we are not using information from the hypothetical data component to consider how policy professionals and researchers update in response to being presented with different kinds of information, even though some participants are randomly provided with confidence intervals and others only point estimates. We focus on each section's intended use because each section of the experiment is best-suited to exploring one topic deeply.

Figure 3: Structure of the Survey



This figure shows the order of the sections in the survey.

particular, that one individual's allocations would be randomly selected and implemented.[12] Participants were then shown real-life data from impact evaluations of cash transfer and school meal programs and asked to allocate funds between these two interventions or to further research. Participants at some workshops were asked to provide their priors and posteriors of the effect of these interventions on school attendance before and after seeing the data and before determining their preferred allocations.[13] The real-life data were taken from AidGrade's data set of impact evaluation results (2016). The point estimates and confidence intervals of the data provided for the cash transfer program were randomized across participants, unbeknownst to the participants. All respondents received the same data for the school meals program. To ensure point estimates and confidence intervals were the only things that varied across treatments, and to ensure there was no deception, minimal detail was provided about the programs. The point estimates and confidence intervals of the different studies shown are reproduced in Table A1.

This approach allows us to see how much differences in treatment effects and confidence intervals affect real-world allocations and, in turn, estimate how much allocations could change if respondents were not biased or if they were to receive information that could help them overcome their biases.

[12] After the end of data collection, this lottery was implemented as described.
[13] There was not enough time to include these questions at all workshops.

4.3 Asymmetric Optimism and Variance Neglect

The basic experimental approach is simple: we elicit participants' priors, randomly show them new information, and then elicit posteriors.

At the conclusion of the section using real data, respondents were told that they were moving on to a new section of the survey and would not need to use any information that was previously provided. In this new section, we asked participants to consider effects of a hypothetical CCT and school meals program. The survey merely asked respondents to "suppose that" they were to guess the effect of a particular program.[14] In the set-up to these questions, respondents were informed that enrollment rates were currently at 90 percentage points, so that respondents could reasonably expect the intervention to improve enrollment rates by at most an additional 10 percentage points. We allowed participants to guess negative values down to -5.

We elicited prior beliefs by allowing participants to place probability weights on intervention impacts in bins (Figure C3). They were then randomized into seeing one of several sets of hypothetical "new data". The respondents were asked to imagine that these data represented replications of studies on the same program. Respondents were not informed that they were presented with randomly-selected results, and audio recordings from surveys where participants consented to being recorded confirm that no individual questioned why they were shown those results in a way that suggests they were aware of the randomization.

Participants were presented with hypothetical results from two studies that either featured a positive or negative outlier relative to their stated prior ("good news" and "bad news") and that either provided no confidence intervals, small confidence intervals or large confidence intervals. The positive or negative outlier was randomly selected but influenced by their stated prior (such that a positive outlier was always two percentage points above the prior mean, and vice versa for a negative outlier). The confidence intervals that were provided were randomized independently from the prior. The "new data" are described in more detail in Table A2, using the example of someone who previously reported they thought enrollment rates increased by 2 percentage points.[15]

[14] No deception was involved.

[15] Since the different randomizations could result in confidence intervals stretching up to 5 above or below the initial mean that respondents provided, we could only follow the above randomization strategy for those who initially stated they expected treatment effects between 0 and 5 percentage points (otherwise, confidence intervals would be cut off in the graphical representation). We believe that this is a reasonable range, especially given that respondents knew that baseline enrollment rates were 90 percentage points. Appendix Table A3 shows that 82.2% of responses indeed fell within this range. To ensure the graphical representations were properly displayed, those who stated expected values lower than 0 were shown the same data as those who stated expected values of 0, and those who stated expected values higher than 5 were shown the same data as those who stated expected values of 5. This poses a slight problem for tests of asymmetric optimism, since people who stated expected values greater than 5 will tend to see new data lower than their priors. If we saw them updating less on these values, we would be unable to attribute that

All study estimates were presented as bar charts, as these represent a common form in which results are displayed to policymakers. Figure C4 provides an example. The order in which the two data points were provided (left to right) was randomized. The data were also described in the text. After viewing these data, participants were again presented with a set of slider bars and asked to put weights on different effect ranges, capturing their posteriors.

By eliciting participants' μ_{t-1}, σ^2_{t-1} and μ_t in this manner, and given that we experimentally provide them with Y_i and σ^2_i, we can calculate k using $\mu_t = \mu_{t-1} + k(Y_i - \mu_{t-1})$.[16] If we observe that $k^+ > k^-$, where k^+ represents the calculated k for those receiving the positive outlier treatments and k^- represents the calculated k for those receiving the negative outlier treatments, that would be evidence of asymmetric optimism.[17] Using the notation of the model, we can estimate γ such that respondents are Bayesian updating with the correct k based on the wrong signal. Note that to test for asymmetric optimism, we do not require normally-distributed priors or posteriors.

We can also compare the responses of those who receive a signal with a small confidence interval and those who receive a signal with a large confidence interval. If we observe that $\partial k^B/\partial\sigma^2_i > \partial k/\partial\sigma^2_i$, where k^B represents the k of a Bayesian updater, that would be evidence of variance neglect.[18] We can then estimate λ.

Since the estimation of λ depends on having normally distributed priors and data, in some analyses we restrict attention to those who report normally distributed priors and posteriors, on the assumption that if they have normally distributed priors and posteriors

to the values being lower than their priors - it could just be that the people who have high initial priors are also less likely to shift their priors. For this reason, for tests of asymmetric optimism we will restrict attention to those whose priors fell between 0 and 5. However, we still use all data for tests of variance neglect.

[16] Our primary focus is on calculating k from the distributions that respondents provided. However, each time we asked for a prior or posterior, we also asked respondents to first provide their best guess of the effect of the program. We use these as estimates of their μ_{t-1} and μ_t in an alternative specification which produces similar results. Results available upon request.

[17] A couple of subtle technical points should be discussed. First, in principle, it is possible that individuals who start with particularly large or small priors may find their distributions partially "cut off", introducing structural correlation between the mean expectation and the variance. In practice, this does not appear to be a problem: we observe only insignificant correlations (a correlation coefficient of 0.002 between the prior mean and variance and a correlation coefficient of 0.074 between the posterior mean and variance). Further, as illustrated by Table A3, very few individuals provide particularly large or small priors. Relatedly, it is possible that individuals have asymmetric priors. We tested for symmetry with two tests (Cabilio and Masaro, 1996; Miao et al., 2006) and found that for the vast majority of the sample symmetry cannot be rejected (95% or 88%, respectively, of those with normally-distributed priors, and 89% and 81%, respectively, unrestricted). These tests still restrict attention to those who put weight in at least 3 bins, since we can't say much about the shape of the distributions if they use fewer bins.

[18] Specifically, recalling that for someone with variance neglect, $k^B_S - k^B_L > k^{VN}_S - k^{VN}_L$, we can rearrange this equation to form $k^B_S - k^{VN}_S > k^B_L - k^{VN}_L$. Then, it is clear that regressing the difference between how a Bayesian would update on certain information and how a respondent updated $(k^B - k)$ on whether or not someone observes large confidence intervals ("Large CI") is a valid test of variance neglect.

they would also have believed the data to be normally distributed.[19] As will be described, nearly 80% of the sample appeared to have normally distributed priors and normally distributed posteriors. Those who have normally-distributed priors and posteriors generally appear to be similar to those who do not on observable characteristics (Table A4).[20]

At the same time, k has an intuitive interpretation even if the priors or posteriors are not normal: it represents the extent to which individuals weigh the new information relative to their priors. Thus, we will also present some results for the full sample, regardless of their prior or posterior belief distributions.[21] In practice, we will observe that the estimates of γ and λ do not greatly depend on the sample used.

4.4 Information Treatments

If policymakers are biased, could some types of information help them make better decisions? To further examine how the statistics provided may affect estimates, we provide respondents with one of several types of information (Section B). These different types of information are provided in the context of an introductory question that asks respondents to estimate temperature. Participants are randomized into receiving point estimates without confidence intervals; point estimates with confidence intervals; point estimates with confidence intervals and the interquartile range; and point estimates with confidence intervals, the interquartile range, and maximum and minimum values. These treatments were constructed so that each subsequent arm contains the same information as the previous arm plus some additional information; in other words, the treatments can be thought of as ordered. Figure C6 illustrates.

4.5 Incentives

Policymakers, policy practitioners and researchers were offered a token gift in the workshops (chocolate or coffee costing approximately $5 USD) in exchange for their time. In addition, participants were informed that at the end of the study, one response would be

[19] It remains possible that some individuals had seemingly normally distributed priors and posteriors without believing the new information represented draws from a normal distribution.

[20] The one exception is that individuals who have larger priors are slightly less likely to have normally-distributed priors. This result is driven by individuals with large priors, who make up a small share of the sample.

[21] It should be noted that while this part of the experiment was framed to respondents as explicitly being about predicting the effects of replications, some individuals may have been thinking about the possibility of there being contextual differences between settings. However, in practice this did not appear to be prevalent. After every quantitative question, we asked a "why" question. Specifically, we asked: "Could you please describe how you determined to put these weights on these ranges?" after respondents provided their prior distributions and "Why do you think this? How did you come up with your new estimate?" after they provided their posteriors. In responses to these questions, extremely few respondents refer in any way to context - an interesting finding in and of itself. Nonetheless, we cannot rule out that some individuals may have been thinking of context without it being reflected in their open-ended responses.

drawn at random and awarded an additional prize: a MacBook. We did not further incentivize responses because we were concerned that policymakers might fear giving a "wrong" answer, so we did not want to increase the salience of the possibility of answering "incorrectly" by offering incentives for "correct" answers. The same incentives were offered to participants at the World Bank and IDB headquarters. For those interviews conducted over videoconference, a $15 Amazon voucher was provided, again without further conditions, along with entry to the MacBook raffle. Enumerators were trained to encourage participants who feared giving a wrong answer that we merely wanted to know what they thought given the information we provided. Despite the lack of formal incentives, respondents appeared to take the survey seriously, with the median respondent spending 32 minutes on it.[22] Respondents also appeared to provide consistent answers, as will be described in the results section.[23] The direct interaction with an enumerator may have contributed to respondents putting effort into their actions.

5 Results

5.1 Descriptive Statistics: Priors and Posteriors

Figure 4 plots the distribution of prior means by subgroup in the policymaker and researcher sample; Figure 5 provides the distribution of individual-level prior standard deviations for policymakers and researchers.[24] Notably, we can already discern several key differences between policymakers, policy practitioners and researchers. First, policymakers had notably higher prior means than researchers ($p<0.0001$).[25] Further, policymakers had significantly narrower priors than researchers ($p<0.001$).[26] This difference in the width of prior beliefs is not a function of knowledge of the interventions included in the experiment; while familiarity with the topics in the experiment was correlated with narrower priors, policymakers reported significantly less familiarity than researchers with the topics in question ($p<0.0001$).[27]

What "should" the priors have been if they were accurate? The question is complicated by the fact that there is no exact real world match for the component using hypothetical data. Nonetheless, we can turn to results from meta-analyses to consider what a reasonable

[22] This excludes those who, in the data, appear to have taken more than an hour on the survey; we assume these represent individuals who started to take the survey, ran out of time before the end of the workshop break, and continued it in the next break in the workshop.

[23] For example, those who provided narrower priors updated less on the new information.

[24] Policy practitioners fall in between and are available upon request.

[25] They also had higher prior means than policy practitioners ($p<0.01$), and policy practitioners also had higher prior means than researchers ($p<0.001$).

[26] Policy practitioners had narrower priors than researchers as well ($p<0.001$). policymakers had insignificantly narrower priors than policy practitioners.

[27] Figure A1 shows distributions of prior means by expertise.

range of values might be. Vivalt (2020) describes a set of point estimates from 36 different studies of the effects of CCTs on enrollment rates, collected in the process of meta-analysis. Some of these point estimates were used to construct the real data component, but we can also consider the full range of estimates. The set of 36 estimates had a mean of 4.4 and a standard deviation of 2.7 when restricting attention to the 0-10 percentage point range that is the most applicable to the hypothetical data component.[28] The full range of observed point estimates is plotted against respondents' prior means in Figure 4.[29]

Policymakers' priors were similarly distributed to the estimates in the CCT data set. However, the standard deviation of the prior means across policymakers is not the same as the standard deviation that describes the narrowness of the average policymaker's distribution of probability weights. The policymakers may have been right to collectively anticipate a broad range of possible effects while being individually overly sure of their own estimates.

Finally, we test for normality of prior distributions using a Kolmogorov-Smirnov test. This test is typically conducted for continuous distributions, rather than for data that fall in discrete bins, as in our case. However, the Kolmogorov-Smirnov test can be applied to discrete data with modifications. It should also be noted that Kolmogorov-Smirnov tests are not well-powered for distributions made up of a few discrete bins, as we have. While 15 bins were available, most respondents' estimates fell into 5 or fewer bins. In 7.2% of priors, respondents put weight in only one or two bins, and we can neither prove nor disprove that the priors are normally distributed. Kolmogorov-Smirnov tests reject an additional 9.3% as non-normal.[30]

Overall, 79% of policymakers, policy practitioners and researchers reported prior and posterior distributions that were in three or more bins and consistent with being normally distributed according to a Kolmogorov-Smirnov test.[31] We will present results both for those who pass or do not pass the Kolmogorov-Smirnov test in alternative specifications. In all

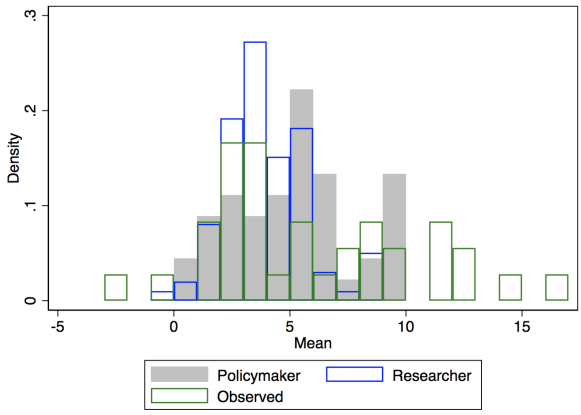
[28] As respondents were told that baseline enrollment rates were already at 90%.

[29] We can't provide similar comparisons for the effects of school meals programs on enrollment rates given there are many fewer studies on this intervention-outcome combination. The same data set has only three estimates for the effects of school meals programs on enrollment rates.

[30] k could not be calculated for 11.8% of observations for the mechanical reason that the point estimate that respondents were shown, which was based on the first mean value that they stated, turned out to be exactly equal to the mean that we calculated from their putting weight in bins. Remember that we asked respondents to first state an integer value and then put weight in bins to make the weighting part of the exercise easier. We take these weights as the most accurate estimate of their prior mean, though we use the integer values in a robustness check. It might seem unusual for the new data to be exactly centered on the midpoint of their prior distribution, but this could happen if a respondent gave a perfectly symmetric distribution.

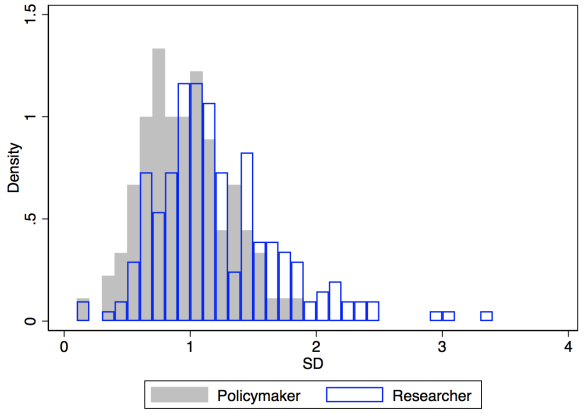
[31] We will implicitly cover the posterior mean in discussing results. The posterior distributions were mostly similar to the distributions of the priors: among those individuals who had priors spanning at least three bins and which could not be rejected as normal by a Kolmogorov-Smirnov test, 1.4% had posterior distributions that fell in 1-2 bins and a further 4.6% of the posterior distributions were rejected as not normal by a Kolmogorov-Smirnov test.

Figure 4: Policymakers Have Larger Priors



This figure plots the distribution of prior means against a distribution of effects from actual studies. As respondents were told baseline rates were 90%, the closest comparison might be to study estimates between 0 and 10. Few responses fall below 0, suggesting respondents are paying attention and averse to giving negative estimates. Most fall between 0 and 5. Results for policy practitioners are omitted for legibility; they fall between those of researchers and policymakers and are available upon request.

Figure 5: Policymakers Have Narrower Priors



This figure plots the distribution of σ_{t-1} across respondents by profession. Results for policy practitioners are omitted for legibility; they fall between those of researchers and policymakers and are available upon request.

cases, however, we will restrict attention to those who place some probability weight in three or more bins. This is 1) so as to have greater comparability between the specifications that depend on normality assumptions and those that do not; 2) because we can more precisely estimate σ for these individuals; and 3) because if someone put all their probability weight in one or two bins they may be less likely to be putting effort into the experiment. Nonetheless, including those who placed weight in only one or two bins does not qualitatively change results, and results with these individuals included are available upon request.[32]

5.2 Descriptive Statistics: Distribution of k

Figure 6 plots histograms of the distribution of k among those with normally distributed priors and posteriors, illustrating a relatively large range, with clusters of estimates around 0 and 1. Note that k should generally fall between 0 and 1, with those who take the data mean as their posterior mean having $k = 1$ and those who stick with their initial mean having $k = 0$. Only 55% of estimates fall within this range, which will be discussed more later.

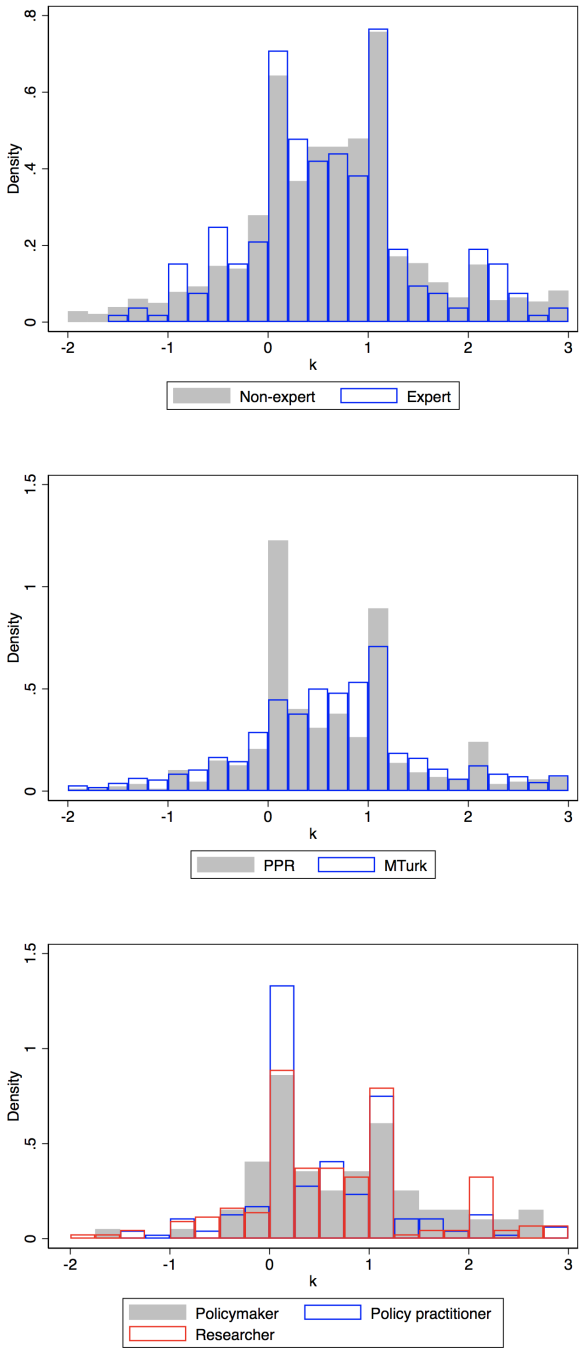
To better understand the data, we break results down by stated familiarity with the types of interventions discussed and we also include comparisons to responses from MTurk workers, who reported less familiarity with the types of interventions discussed. Figure 6 distinguishes between responses to a "knowledge" question asked of all respondents: for each intervention, respondents were asked to specify whether they had "never heard of it", "heard of it but never heard of any studies on it", "heard of it and heard of some studies", or "heard of it and very familiar with studies". Those who reported greater familiarity with a type of intervention typically updated less in response to new information, as expected. Similarly, the middle plot in Figure 6 shows that MTurk workers updated more on the data than policymakers, policy practitioners or researchers. This makes intuitive sense given their lesser familiarity with the interventions. Policymakers, who stated they were less familiar with studies than policy practitioners or researchers, also had higher values of k.[33]

5.3 Tests for Biases

The wide dispersion of k values complicates testing for biases. Values of k within the -0.5 to 1.5 range could simply reflect noise in using the slider bars. However, we may think that values of k that are much smaller or larger suggest that individuals are using a different

[32] A couple of results are made slightly more statistically significant when these individuals are included.
[33] Corresponding figures for $k^B - k$ are included in the Appendix as Figure A2.

Figure 6: Distribution of k



This figure plots values of k calculated from respondents' reported μ_{t-1}, μ_t', and the provided Y_i values. Values below -2 or above 3 are not included for legibility. The top plot distinguishes between responses to a "knowledge" question asked of all respondents: for each intervention, respondents were asked to specify whether they had "never heard of it", "heard of it but never heard of any studies on it", "heard of it and heard of some studies", or "heard of it and very familiar with studies". For visual clarity, the first three categories are collapsed into "Non-expert" and the last considered "Expert".

updating heuristic or misinterpreting the exercise.[34] We have limited information on the heuristics that respondents are using, based on descriptive responses to questions about why they answered the way they did.[35] As we cannot fully identify the reasons why some responses imply particularly large or small values of k, we separate individuals into different groups: those with k in the expected $0 \leqslant k \leqslant 1$ range; those with $-0.5 \leqslant k \leqslant 1.5$, where large or small values could represent noise; and the full sample. We will present results for each group.

Table 3 presents results for regressions of respondents' calculated values of k on whether they received the "good news" treatment. Robust standard errors are used, clustering observations at the individual level. Receiving the good news treatment significantly affected k for most specifications, whether or not we restrict attention to those with normally-distributed priors and posteriors. The first and fourth column present results of a logistic regression using the full sample, while the other columns present OLS results restricting k to various ranges.

Recall that to test for variance neglect, we need to consider what someone who was Bayesian updating would do. Thus, in Table 4, we construct $k^B - k$, where k^B is the value that k should have taken if respondents were Bayesian given their stated priors and the Y_i and σ_i^2 they observed. The model implied that $k_S^B - k_L^B > k_S^{VN} - k_L^{VN}$; we test this by regressing $k^B - k$ on whether the respondent saw large or small confidence intervals. In Table 4, we observe that $k_S^B - k_S$ is indeed sometimes greater than $k_L^B - k_L$, as predicted, though this estimate is noisier and often not significant. All regressions are restricted to those with normally-distributed priors and posteriors for the simple reason that we cannot calculate k^B for those with non-normal priors and posteriors without additional assumptions.[36]

[34] For example, k could be greater than 1 if respondents have in mind a certain prior based on older information they hold, are surprised by the results they are shown, and infer there is a time trend that explains the discrepancy between their older priors and the new information they were shown. They may then expect that the effects captured in a subsequent evaluation would also be subject to the same time trend.

[35] In particular, after every question, respondents were asked to describe why they answered the way they did, the enumerator summarized this response in a text box, and respondents were given an opportunity to amend the summary. For a subset of responses, the full audio transcripts of this exchange were also captured.

[36] Recall that in order to calculate k^B we need to know individuals' priors as well as what they think the distribution of the new data is. For those with normally-distributed priors and posteriors, it is reasonable to assume they also believe they face normally-distributed new data. While we could calculate k^B by simulation for those without normally-distributed priors or posteriors, it is less clear what we would assume individuals believe about the distribution of the new data.

Table 3: Tests of Asymmetric Optimism: Regressing k

	Normally-distributed			Any distribution		
	$I(k > 0.5)$	k		$I(k > 0.5)$	k	
	(1)	(2)	(3)	(4)	(5)	(6)
Good News	2.107***	0.076	0.172***	2.053***	0.053	0.174***
	(0.42)	(0.06)	(0.06)	(0.39)	(0.05)	(0.06)
Observations	429	229	287	475	254	317
k restrictions	-	0-1	-0.5-1.5	-	0-1	-0.5-1.5

This table reports the results of regressions of k on an indicator of whether the respondent received "good news", *i.e.* positive new data relative to their priors. Columns (1) - (3) report results using the sample of policymakers, policy practitioners and researchers with normally-distributed priors and posteriors; Columns (4) - (6) report results using the sample of policymakers, policy practitioners and researchers regardless of their distributions of priors and posteriors. Columns (1) and (4) present the results of a logistic regression of $I(k > 0.5)$ on whether the respondent received good news, using exponentiated coefficients; Columns (2) and (5) regress k on whether the respondent received good news, restricting attention to only those observations for which $0 \leqslant k \leqslant 1$; Columns (3) and (6) also regress k on whether the respondent received good news, considering only those observations for which $-0.5 \leqslant k \leqslant 1.5$. Only those who provided prior means between 0-5 percentage points were included in the tests for asymmetric optimism, as we were unable to show new data above or below higher or lower priors without going out of range of what could be displayed. Including those outside of this range could introduce bias. Nonetheless, results are comparable when those outside this range are included, and those results are available upon request.

Table 4: Tests of Variance Neglect: Regressing $k^B - k$

	$I(k^B - k > 0)$	$k^B - k$	
	(1)	(2)	(3)
Large C.I.	1.342	-0.118**	-0.073
	(0.29)	(0.06)	(0.06)
Observations	349	206	256
k restrictions	-	0-1	-0.5-1.5

This table reports the results of regressions of $k^B - k$ on an indicator of whether the respondent saw large confidence intervals as opposed to small confidence intervals. Respondents can be included regardless of their priors, unlike in testing for asymmetric optimism, but cases in which respondents were randomized into seeing no confidence intervals are excluded. Columns (1) - (3) report results using the policymakers, policy practitioners and researchers sample. Column (1) presents the results of a logistic regression of $I(k^B - k > 0)$ on whether the respondent saw a large confidence interval as opposed to a small confidence interval, using exponentiated coefficients; Column (2) regresses $k^B - k$ on whether the respondent received a large confidence interval as opposed to a small confidence interval, restricting attention to only those observations for which $0 \leqslant k \leqslant 1$; Column (3) also regresses $k^B - k$ on whether the respondent received a large confidence interval or a small confidence interval, considering only those observations for which $-0.5 \leqslant k \leqslant 1.5$. Again, cluster-robust standard errors are used. All results are on the sample of those who have normally distributed priors and posteriors, as k^B is calculated under that assumption.

5.4 Heterogeneity by Profession and Subject Matter Expertise

We consider heterogeneity by profession in Table 5. Depending on the specification, policy practitioners and researchers shift less towards the new data's point estimate than MTurk workers, while policymakers behave similarly to MTurk workers. This would be consistent with a story in which researchers or policy practitioners had more background knowledge or narrower priors (as they do in fact have), and it does not require a difference in updating. Policymakers, policy practitioners and researchers do not appear to experience significantly more or less asymmetric optimism than MTurk workers, the subgroup left out. We also cannot reject that they were equally insensitive to confidence intervals, as indicated by the regression coefficients on the interaction of being a policymaker, policy practitioner or researcher and large confidence intervals in Columns (4)-(6).

We also explore the role of specific subject matter expertise. In particular, a diverse range of policymakers, policy professionals and researchers attended the workshops, and it is plausible that individuals who have more experience with the conditional cash transfer and school meals programs could update differently. We use the self-reported familiarity questions for these tests. Recalling that the questions about familiarity asked whether, for a given intervention, a respondent had "never heard of it", "heard of it but never heard of

Table 5: Heterogeneity by Profession

	$I(k > 0.5)$	k		$I(k > 0.5)$	k	
	(1)	(2)	(3)	(4)	(5)	(6)
Good News	1.766***	0.091***	0.118***			
	(0.23)	(0.03)	(0.04)			
Policymaker	1.277	-0.033	0.060	0.935	0.103	0.064
	(0.43)	(0.09)	(0.09)	(0.28)	(0.09)	(0.10)
Policymaker *	1.025	-0.120	-0.027			
Good News	(0.52)	(0.15)	(0.16)			
Pol. Pract	0.540***	-0.163***	-0.164***	1.291	0.088	0.053
	(0.12)	(0.05)	(0.05)	(0.33)	(0.07)	(0.07)
Pol. Pract *	1.217	-0.030	0.017			
Good News	(0.41)	(0.09)	(0.10)			
Researcher	0.811	-0.128**	-0.160***	1.068	0.097	0.013
	(0.18)	(0.06)	(0.06)	0.29	(0.07)	(0.08)
Researcher *	1.271	0.027	0.114			
Good News	(0.42)	(0.10)	(0.10)			
Large C.I.				0.770**	-0.121***	-0.156***
				(0.10)	(0.03)	(0.04)
Policymaker *				1.967	-0.071	-0.067
Large C.I.				(0.89)	(0.14)	(0.13)
Pol. Pract *				1.642	0.036	0.121
Large C.I.				(0.60)	(0.09)	(0.10)
Researcher *				1.714	-0.016	0.103
Large C.I.				(0.62)	(0.11)	(0.11)
Observations	1482	645	881	1374	660	895
k restrictions	-	0-1	-0.5-1.5	-	0-1	-0.5-1.5

This table considers heterogeneity by profession, including both the policymaker, policy practitioner and researcher sample and MTurk sample and interacting indicator variables for each subgroup. Columns (1) and (4) present (alternatively) the results of a logistic regression of $I(k > 0.5)$ on whether the respondent received good news or a logistic regression of $I(k^B - k > 0)$ on whether the respondent saw large confidence intervals, using exponentiated coefficients; Columns (2) and (5) present results of regressions of k and $k^B - k$, respectively, on whether the respondent saw large confidence intervals, restricting attention to those observations for which $0 \leqslant k \leqslant 1$; Columns (3) and (6) consider only those observations for which $-0.5 \leqslant k \leqslant 1.5$. The tests for asymmetric optimism and variance neglect were conducted on slightly different samples. The tests of asymmetric optimism require respondents to have mean priors between 0 and 5, while the tests for variance neglect require that respondents be randomized into seeing small or large confidence intervals (as opposed to no confidence intervals). All regressions restrict attention to those with normally-distributed priors and posteriors.

any studies on it", "heard of it and heard of some studies", or "heard of it and very familiar with studies", we consider the final category as indicating specific expertise. Tables A5-A6 report the results of regressions including these interactions. We observe that asymmetric optimism does not appear to depend on the degree of expertise, but variance neglect does at least in one specification. This latter result, however, is only suggestive, because those who have more specific expertise tend to have narrower priors and it may be more difficult to observe variance neglect among those with narrower priors.[37]

5.5 Estimating γ and λ and Simulations

We can obtain the value of γ and λ to test for the presence of asymmetric optimism and variance neglect respectively for each individual in the study. Both γ and λ vary substantially across individuals, and the median may be a more informative summary statistic than the mean.[38] Table 6 summarizes the different parameter values estimated for each subgroup. We find a median value of γ of 0.23 and a median value of λ of -0.99 in the main policymakers, policy practitioners and researchers sample. If γ is positive, it indicates asymmetric optimism; if λ is negative, it indicates that participants update too much in response to the new information, given its confidence interval. Note, however, that λ is sensitive to the width of individuals' priors. The intuition is that if someone had more uncertainty in their prior beliefs, they would update more differently on small versus large confidence intervals than if they had narrower priors. This means it is easier to observe large deviations in updating from Bayesian updating for people who have wider prior distributions than those who have narrower prior distributions. This will be relevant when we consider differences in λ by subgroup because the different subgroups have different prior distributions. In particular, recall that researchers had relatively wide priors, while policy practitioners had relatively narrow priors, with policymakers falling in between.[39]

We use quantile regression to test whether each subgroup's median is significantly different from 0 and from each other. Almost all estimates of γ and λ for each subgroup are significantly different from 0, except for researchers' γ being insignificantly different from 0 in both specifications and policy practitioners' γ being insignificantly different from 0 in one specification. However, we do not observe large differences between the groups. Policy practitioners and researchers have statistically different λ, but only when the prior variance is not controlled for. Controlling for the prior variance, the only marginally statistically

[37] This is a result of Bayesians not updating much if they have narrower priors.

[38] It should be noted that these statistics are calculated with a numerator and denominator term, and the parameter estimates are hence sensitive to small values in the denominator.

[39] In principle, it theoretically possible that γ could vary with prior means, such as if those who initially skeptical that an intervention would work display less asymmetric optimism. Empirically, however, we observe no relationship between γ and individuals' prior means.

significant difference is between the λ of policy practitioners and MTurk workers ($p<0.1$), and no subgroup has a significantly different γ from any other subgroup.

This is consistent with each group being affected by these biases, but there not being significant differences in biases across subgroups, as we saw in the earlier regressions (Table 5). It is possible that some differences exist but that these are too small for us to be powered to detect them.

Table 6: Median Estimated Parameter Values

	Normally-distributed		Any distribution	
	γ	λ	γ	λ
Policymakers	0.52	-0.89	0.51	-0.86
	(0.055)	(0.001)	(0.056)	(0.000)
Policy practitioners	0.22	-0.66	0.34	-0.66
	(0.134)	(0.000)	(0.006)	(0.000)
Researchers	0.18	-1.17	0.20	-1.06
	(0.238)	(0.000)	(0.161)	(0.000)
PPR	0.23	-0.99	0.33	-0.91
	(0.021)	(0.000)	(0.000)	(0.000)
MTurk	0.29	-1.17	0.31	-1.17
	(0.000)	(0.000)	(0.000)	(0.000)

This table reports the median values of γ and λ for each type of respondent. "PPR" represents the pooled sample of policymakers, policy practitioners and researchers. Columns (1) and (2) show the median values for those with normally-distributed priors and posteriors, while columns (3) and (4) show values without these restrictions. p-values are provided in parentheses below each result, representing a test of the hypothesis that the median is significantly different from 0.

What are the practical implications of these results? Figure A3 presents some simulations showing how individuals with different biases might update on the same evidence base, relative to a Bayesian. For these simulations, we assume new study results arrive as draws from a normal distribution. This normal distribution is selected to match the real evidence base on cash transfer programs as closely as possible.[40] For priors, we use the real priors observed.

Precisely how the biases affect beliefs depends on (i) the initial beliefs, (ii) how many new studies there are, and (iii) what they find. The literature suggests that typically there may only be six papers on a topic in development economics, conditional on there being more than one paper on the topic.[41] If we consider how the typical respondent might update on six new studies, based on our results they would end up believing the

[40] These are the same study results as in Figure 4.

[41] Vivalt (2020). Many studies do not overlap at all in terms of outcomes studied.

intervention is approximately 9% more effective than a Bayesian would believe it to be (with a point estimate of 4.5 rather than 4.1 percentage points), and the variance of their posterior beliefs would be 31% smaller (generating a 95% credible interval that is 2.2 as opposed to 3.2 percentage points wide). We do not want to lean too hard on these numbers, given they will vary according to the inputs. However, we can observe that if the model is correct the different behavioral biases are not fully "self-correcting" over time but persist even as the number of studies approaches 50, representing a very large number of studies on a given intervention in development economics.[42]

Table 6 also highlights that there are generally minimal differences between those with normally-distributed priors and posteriors and those with any distribution, as we might expect when the former comprises about 80% of the latter.

5.6 Changes in Allocations

In this section, we present reduced-form estimates of the impact of seeing more positive results data on allocations. In some workshops, we did not ask this part of the survey due to time constraints, so we only have allocations from 308 individuals.

There are several reasons we may expect that information provided may not influence participants' allocations. First, in our experiment participants only receive information about a particular outcome variable, and they may care about different outcomes. Second, they may not feel like they have sufficient information to materially change their beliefs.[43]

The average allocations to cash transfer programs, school meals programs and further research, respectively, were 36.4%, 34.3% and 29.3%. Recall that in this part of the survey, respondents randomly viewed a selection of real data on cash transfer programs, with point estimates of 1 and 4 or point estimates of 2 and 5 and with confidence intervals ranging 2 or 5 above and below those values. Viewing the larger point estimates resulted in an increase in the amount allocated to cash transfers of 7.1 percentage points; viewing results with larger confidence intervals resulted in a decrease in the amount allocated to cash transfers of 8.4 percentage points. Interestingly, when respondents saw large confidence intervals, they allocated 5.7 percentage points more to further research. Seeing a large confidence interval did not temper the response to large point estimates. Results are presented in

[42]The variance neglect curve should eventually approach the Bayesian curve. While the curves are governed wholly by the model, they are fit with the parameter values estimated and the figure demonstrates they need not quickly converge.

[43]As the section that asked participants to make allocations used real rather than hypothetical data, no study details could be provided, lest they update on characteristics that varied between studies, as previously explained. Updating based on study characteristics would be more realistic but would prevent us from identifying the effect of observing more positive results. A companion paper (Vivalt et al., 2021) examines how a similar sample values different studies based on their characteristics, using a discrete choice experiment.

Table 7.

Table 7: Regressions of Allocations on Evidence Shown

	CCT Funding			Research Funding		
	(1)	(2)	(3)	(4)	(5)	(6)
Large Point	7.137***		7.450***	1.910		-0.347
Estimate	(1.95)		(2.68)	(2.48)		(3.30)
Large Confidence		-8.379***	-8.291***		5.741**	3.132
Interval		(1.98)	(2.58)		(2.48)	(3.60)
Large Point			-0.354			4.435
Estimate * Large C.I.			(3.78)			(4.94)
Observations	308	308	308	308	308	308

This table reports the results of regressions of allocations to CCTs or to further research on whether or not large or small point estimates were shown (a mean difference in the point estimates of 1 percentage point), whether large or small confidence intervals were provided ("large" confidence intervals extended 5 percentage points above or below the point estimate; "small" confidence intervals extended 2 percentage points above or below the point estimate), and the interaction of the two. We were unable to ask the allocation question at all workshops due to time constraints, hence the smaller sample size. Robust standard errors are used.

It is instructive to consider how much allocations would change if respondents were Bayesian updating. Recall that we estimated the median γ to be equal to 0.23 and the median λ to be equal to -0.99 overall, or γ=0.52 and λ=-0.89 among policymakers. If a policymaker suffering from optimism misperceived a signal that they thought a certain option had a point estimate that was 0.52 higher than a Bayesian might think it to be, the estimates in Table 7 would suggest they might allocate 3.7 percentage points more than a Bayesian would to that option, or 10.2% more than the average allocation to cash transfers of 36.4 percentage points. Similarly, Table 7 suggests respondents dislike noisy data and would allocate 8.4 percentage points less to data with larger confidence intervals, yet given that they also misinterpret the large confidence intervals as smaller than they actually are, they would allocate 14.5 percentage points less to this intervention if they interpreted the confidence intervals correctly. These back-of-the-envelope calculations rely on the specific functional form and values considered so we do not want to lean too hard on them.[44] Nonetheless, they suggest the biases can have meaningful effects.

How would these biases affect allocations in equilibrium? We might imagine that if owing to their biases a policymaker wanted to allocate more to every program, they would

[44]Specifically, here we are converting the confidence intervals from Table 7 into variances and generating a distaste per unit variance, multiplied by λ. But there is no reason to think the preference for more precise estimates is linear in this way.

not have enough money to do so, and so we might imagine they would still make the same allocations across programs that they would if they were Bayesian. However, this ignores some important factors. First, policymakers are unlikely to receive information for many projects at the same time. For example, a policymaker could have priors about a default option and only observe a signal about one other program that they are particularly excited about and want researched. If that signal is selected to be particularly positive, as sometimes occurs (e.g., Allcott (2015)), that would lead policymakers to update even more towards their preferred option when it may not actually be the best choice. Further, if some evaluations find larger point estimates than others due to having been imprecisely estimated, the observed asymmetry in updating would bias policy towards these types of programs. Smaller, NGO-implemented pilot programs often have larger effects and scale-ups often fare worse (Bold et al., 2018; Vivalt, 2020); the biases we observe would nudge policymakers towards overoptimism and overconfidence in these small studies' results. Separately, errors in updating could lead policymakers to not achieve their desired goals. For example, if a government were aiming to reduce greenhouse gas emissions by a certain percent and were overoptimistic about the effects of programs to reduce it, they would not hit their targets.

Variance neglect could also have a cumulative effect. In the extreme case, after repeated instances of overupdating based on initial noisy data, the policymakers would have very narrow posterior beliefs and be completely unresponsive to any future study results. This is particularly concerning when larger studies often follow more targeted pilots; by the time of the larger studies, beliefs may be relatively fixed.

5.7 How Much Information Should We Provide?

The type of information that was provided affected belief updating. Table 8 shows regression results from Section B in the survey. Policymakers, policy practitioners and researchers update more in response to more information, with the main difference being between seeing point estimates and/or confidence intervals as opposed to seeing point estimates, confidence intervals, and the interquartile range, with or without maximum and minimum values. The significant results imply that some of the previously observed biases may be mitigated by strategic provision of information. To be clear, we are not claiming that respondents *should* have updated the same way on these different pieces of information - they are different statistics, providing different information. Rather, these results simply suggest a lever that could be used in situations in which one worries that an audience will not be very responsive to new research findings.

Table 8: Impact of More Information

	$I(k > 0.5)$	k	
	(1)	(2)	(3)
Point Estimate	0.446*	-0.133***	-0.091*
	(0.19)	(0.05)	(0.05)
Point Estimate and	0.397**	-0.103*	-0.093*
Confidence Interval	(0.17)	(0.05)	(0.05)
Point Estimate,	0.793	-0.021	-0.019
Confidence Interval, IQR	(0.37)	(0.04)	(0.04)
Observations	388	314	373
k restrictions	-	0-1	-0.5-1.5

This table shows the impact of providing more information on k on the policymakers, practitioners and researchers sample. Each type of information is represented by an indicator variable, with the category left out being shown point estimates, confidence intervals, the interquartile range, and maximum and minimum values. Column (1) shows the results of a logistic regression on $I(k > 0.5)$ with exponentiated coefficients. Column (2) restricts attention to $0 \leqslant k \leqslant 1$, and Column (3) restricts attention to $-0.5 \leqslant k \leqslant 1.5$.

5.8 Causes of Variance Neglect

So far, it is not clear what is driving the observed variance neglect. Respondents could be misinterpreting confidence intervals, or they could be interpreting them correctly but not putting weight on them for other reasons (*e.g.* inattentiveness). To shed light on this issue, we run a set of additional experiments with 877 participants on MTurk.[45] In the workshop setting, we could not conduct experiments that looked like numerical exercises, but we can ask these types of questions of MTurk respondents.

First, we check whether respondents still seem to exhibit variance neglect when faced with incentivized questions in a value of information game. Respondents are first asked to provide their full distribution of priors regarding the effect of a particular program on a particular outcome. Then they are given 100 tokens and asked their willingness to pay for information from a replication, after which they can modify their guess. Those who guess correctly within a range will win a bonus. Respondents are asked their willingness to pay for results with various levels of precision, with the understanding that after they provide all these estimates one of the estimates will be selected and the tokens bid on that question taken from them. The more tokens they bid to receive information, the more likely they would be to receive the information, the "cost" of the information being set at some threshold unknown to them. Remaining tokens that are unspent are also worth a small

[45] This sample is distinct from the main experiment's MTurk sample, further described in the appendix.

bonus, so this exercise is incentive-compatible.

We then compare respondents' willingness to pay at each level of precision with a Bayesian's willingness to pay, given the same priors. We pre-specified that those who were unwilling to pay for any of the information sets would be excluded. We might expect that for certain levels of precision, a Bayesian would be willing to pay more for a study's results than our respondents, but for other levels of precision, a Bayesian would be willing to pay less for a study's results than our respondents. In particular, for someone exhibiting variance neglect, they should be willing to pay more than a Bayesian for results with low degrees of precision and less than a Bayesian for results of high degrees of precision.

We also ask respondents to provide their willingness-to-pay for two kinds of confidence intervals - 95% confidence intervals and the equivalent 99% confidence intervals based on the same mean and standard error. This allows us to test whether respondents seem to understand the difference between 95% and 99% confidence intervals by comparing willingness to pay measures for each 95% confidence interval with their corresponding 99% confidence interval. Sample sizes may be more intuitive to understand than confidence intervals, so we also elicit willingness-to-pay for results with equivalent sample sizes. The options that individuals view in this experiment are constructed such that the sample sizes provide the same level of precision as each of the 95% and 99% confidence intervals.

To guard against inattentiveness, we impose a requirement that respondents must successfully pass at least 15 out of 16 attention checks.[46] The attention checks are important because some recent work suggests that the quality of MTurk responses has deteriorated over time (Chmielewski and Kucker, 2019).[47]

The theoretical willingness-to-pay that respondents should have had given their priors if they were Bayesian and their observed willingness-to-pay for results with various confidence intervals and sample sizes is displayed in Figure A4. Consistent with variance neglect, study participants are less responsive to differences in confidence interval widths and sample sizes than someone with the same priors who was Bayesian updating would be; this figure nicely

[46] This was pre-specified. The attention checks were as follows. First, some questions allowed respondents to provide a lower bound that was higher than the upper bound they provided when combining confidence intervals. Second, some questions allowed respondents to provide a lower bound that was higher than the upper bound they provided when creating different confidence intervals, given a 95% confidence interval. Third, some questions allowed respondents to provide answers that were outside the range of possible values stated in the question. Finally, some questions allowed respondents to provide larger confidence intervals than the 95% confidence interval when being asked to provide smaller (*e.g.* 90%) confidence intervals. We count each instance of one of these types of errors as a mistake.

[47] The results we present for the value of information game in Table 9 are based on running the experiment in November-December, 2022. At that time, 90% of our MTurk workers failed the screening questions; we are reasonably confident that the questions are sufficiently challenging and unique to screen out bots. However, for supporting evidence, we ran the same experiment without the sample size questions in 2016, when Chmielewski and Kucker (2019) find minimal evidence of inconsistent answers or bots. We use the exact same screening criteria, and the results are nearly exactly replicated. Results available upon request.

mirrors Figure 1.

Interestingly, respondents were willing to pay more for results that provided a 99% confidence interval than for results with the equivalent 95% confidence interval. For example, a 95% confidence interval that is 5.26 percentage points wide is equivalent to a 99% confidence interval that is 6.92 percentage points wide. Yet, the average respondent was willing to pay 22.43 tokens for results with a 95% confidence interval that was 5.26 percentage points wide and willing to pay 24.02 tokens for results with a 99% confidence interval that was 6.92 percentage points wide. This statistically significant discrepancy points to a fundamental misunderstanding of confidence intervals. For each 95% confidence interval, respondents were willing to pay more for the equivalent 99% confidence interval (Table 9).

Respondents were also more responsive to information about sample sizes than they were to the equivalent confidence intervals. Here, they were generally willing-to-pay more for results with large sample sizes and willing-to-pay less for results with small sample sizes than they were for the equivalent confidence intervals. While they were still far less responsive to differences in sample sizes than a Bayesian would have been, they behaved more like a Bayesian when considering results with different sample sizes than they did when considering results with different confidence intervals.

Table 9: Willingness-to-Pay for Equivalent 95% and 99% Confidence Intervals and Sample Sizes

	95% Confidence Intervals						
WTP	5.26	6.77	8.28	9.78	11.29	12.80	14.31
Bayesian	39.80	25.99	17.67	12.44	8.96	6.58	4.91
95% Interval	22.43	21.32	19.52	18.84	17.96	17.85	17.39
99% Interval	24.02	22.97	20.88	19.89	19.88	18.82	18.49
	(0.0104)	(0.0080)	(0.0268)	(0.0870)	(0.0010)	(0.0727)	(0.0246)
Sample Size	25.87	24.16	20.64	18.56	16.48	15.11	14.61
	(0.0000)	(0.0002)	(0.1153)	(0.6951)	(0.0419)	(0.0002)	(0.0002)

This table shows how the MTurk sample was willing to pay more for every 99% confidence interval than the equivalent 95% confidence interval and behaved more like a Bayesian would when the information about precision was presented in terms of sample sizes rather than confidence intervals. In particular, respondents were willing to pay more for precise results when they were informed about the sample size than when they were informed about the confidence intervals, and they were willing to pay less for imprecise results when they were informed about the sample size than when they were informed about the confidence intervals. Combined with overall low responsiveness to information about precision, this means that providing information about sample size nudged individuals to behave closer to a Bayesian. The numbers in parentheses represent p-values from t-tests of the difference between individuals' willingness-to-pay for 95% and 99% intervals or 95% intervals and sample sizes. Prices are in terms of number of tokens.

Overall, this experiment shows not only that variance neglect may be a quite general

behavioral bias, but also that updating may be improved by strategic choice of how that information is presented. Future work could further explore how providing different statistics can improve updating.

6 Conclusion

The "credibility revolution" has resulted in a significant increase in impact evaluations over the past 20 years (Cameron et al., 2016). More recently, this has been complemented with studies showing that providing training (Mehmood et al., 2021; Toma and Bell, 2023) and new evidence on the effectiveness of specific interventions (Hjort et al., 2021) can both be valued by policymakers and influence their beliefs and decisions. However, if policymakers are biased in the way they interpret this evidence and update their beliefs, it is not guaranteed that the generation of and exposure to new evidence will result in better policy decisions.

We explore this topic by asking how much policymakers' belief updating process differs from Bayesian updating and if there are ways to present evidence that can mitigate these biases. In particular, given that policymakers may see a range of different estimates, how do they consider good news relative to bad news or results with different confidence intervals? These are particularly important questions given well-known concerns with the development economics literature: publication bias (Brodeur et al., 2016), under-powered studies (Ioannidis et al., 2017) and a skewed focus on "easy-to-implement" or small-scale pilot interventions (Allcott, 2015; Bold et al., 2018; Usmani et al., 2022). All of these factors have the potential to artificially inflate the impacts of programs, often with imprecise estimates. When "good news" is favored over "bad news" (asymmetric optimism) and the uncertainty of results is not adequately accounted for (variance neglect), new evidence could increase, rather than reduce the wedge between beliefs and the evidence base.

First, we find that policymakers, policy practitioners and researchers exhibit large differences in prior beliefs. In particular, policymakers are more optimistic about impacts than policy practitioners, who are in turn more optimistic than researchers. Policymakers also have the narrowest priors, despite expressing less familiarity than policy practitioners or researchers with the development programs that were the focus of the study.

Second, we show that, while policymakers suffer from both the biases of interest, this cannot explain the prior differences in beliefs since policymakers, policy practitioners and researchers all exhibit similar biases. These biases do, however, imply that policy professionals and researchers alike may update too much on imprecisely-estimated, inflated estimates and less on more precisely-estimated, modest estimates than a Bayesian would. By parameterizing these biases in a simple updating framework we show that they can lead

to meaningful and persistent differences between beliefs and what the evidence base shows.

What are the implications for how research findings are communicated? First, researchers should put effort into designing larger, better-powered studies to ensure consumers of the evidence are not misled by noisy data, particularly in the presence of publication bias. Beyond ensuring that higher-quality evidence is produced, how the evidence is presented matters. We found that the provision of more detailed summary statistics leads to increased updating. This implies that in cases in which one has to share bad news, providing more information may help. Finally, in a supporting experiment with MTurk workers, we found evidence of respondents being willing to pay different amounts for confidence intervals and sample sizes that conveyed the same information about precision. This suggests an intriguing possibility: that strategic choice of unconventional confidence intervals (such as 90% or 99% confidence intervals) or the presentation of sample sizes may encourage better updating. We hope this work inspires new ideas as to how research results might best be presented and disseminated to policymakers, an under-explored research area in which further innovation could have large benefits.

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.

Angrist, J. D. and J.-S. Pischke (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives 24*(2), 3–30.

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

Bold, T., M. Kimenyi, G. Mwabu, J. Sandefur, et al. (2018). Experimental evidence on scaling up education reforms in kenya. *Journal of Public Economics 168*, 1–20.

Brodeur, A., M. Lé, M. Sangnier, and Y. Zylberberg (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics 8*(1), 1–32.

Cabilio, P. and J. Masaro (1996). A simple test of symmetry about an unknown median. *The Canadian Journal of Statistics 24*, 349–361.

Camerer, C., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai1, S. Isaksson, G. Nave1, T. Pfeiffer, M. Razen, and H. Wu (2016). Evaluating replicability of laboratory experiments in economics. *Science 351*(6280), 1433–1436.

Camerer, C., G. Loewenstein, and M. Rabin (2003). *Advances in Behavioral Economics*, Volume Roundtable series in behavioral economics, Chapter Behavioural Economics - Past, Present & Future. New York: Princeton: Princeton University Press.

Cameron, D. B., A. Mishra, and A. N. Brown (2016). The growth of impact evaluation for international development: how much have we learned? *Journal of Development Effectiveness 8*(1), 1–21.

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

Chmielewski, M. and S. Kucker (2019). An mturk crisis? shifts in data quality and the impact on study results. *Social Psychological and Personality Science 11*.

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

Coville, A. and E. Vivalt (2017). Using subjective expectations to assess research credibility. *Working Paper*.

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.

Delavande, A., X. Gine, and D. McKenzie (2011). Measuring subjective expectations in developing countries: A critical review and new evidence. *Journal of Development Economics 94*, 151–163.

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*.

Duflo, E. and A. Banerjee (2011). *Poor economics*, Volume 619. PublicAffairs.

Easterly, W. (2002). The cartel of good intentions: The problem of bureaucracy in foreign aid. *The Journal of Policy Reform 5*(4), 223–250.

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.

Gelman, A. and J. Carlin (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science 9*(6), 641–651.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, and A. Vehtari (2013). *Bayesian Data Analysis* (3 ed.). Taylor & Francis Ltd.

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 (2021). How research affects policy: Experimental evidence from 2,150 brazilian municipalities. *American Economic Review 111*(5), 1442–80.

Ioannidis, J. P. A., T. D. Stanley, and H. Doucouliagos (2017). The Power of Bias in Economics Research. *Economic Journal*.

Irwin, F. W. (1953). Stated Expectations as Functions of Probability and Desirability of Outcomes. *Journal of Personality 21*(3), 329–335.

Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *American Economic Review 93*(5), 1449–1475.

Kahneman, D. and A. Tversky (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica 47*(2), 263.

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

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

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

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

Malmendier, U. and G. Tate (2005). CEO Overconfidence and Corporate Investment. *The Journal of Finance 60*(6), 2661–2700.

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.

Mehmood, S., S. Naseer, and D. L. Chen (2021). Training policymakers in econometrics. *NBER working paper*.

Miao, W., Y. R. Gel, and J. L. Gastwirth (2006). A new test of symmetry about an unknown median. *Random Walk, Sequential Analysis and Related Topics*, 199–214.

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

Moore, D. and P. Healy (2018). The Trouble with Overconfidence. *Psychological Review 115*(2), 502–517.

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

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

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

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

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

Rabin, M. and D. Vayanos (2010). The Gambler's and Hot-Hand Fallacies: Theory and Applications. *Review of Economic Studies 77*(2), 730–778.

Rogger, D. and R. Somani (2023). Hierarchy and information. *Journal of Public Economics 219*, 104–23.

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.

Toma, M. and E. Bell (2023). Understanding and increasing policymakers' sensitivity to program impact. *Working Paper*.

Usmani, F., M. Jeuland, and S. K. Pattanayak (2022). Ngos and the effectiveness of interventions. *The Review of Economics and Statistics*.

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

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

Vivalt, E., A. Coville, and S. KC (2021). Weighing results: Which attributes matter? *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.

For Online Publication: Appendices

A Additional Figures and Tables

Table A1: Real Data Shown

	Cash Transfers		School Meals	
Intervention	Point Estimates	C.I. Widths	Point Estimates	C.I. Widths
Treatment 1	1 and 4	2	3	4
Treatment 2	1 and 4	5	3	4
Treatment 3	2 and 5	2	3	4
Treatment 4	2 and 5	5	3	4

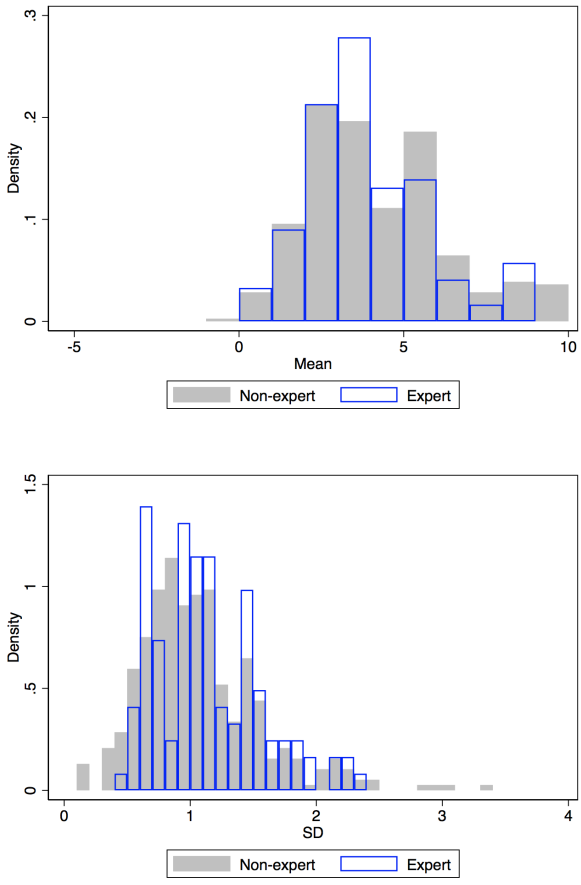
This table shows the point estimates and confidence interval widths that respondents were shown. Each participant saw the school meals result and one randomly-selected result for cash transfer programs.

Table A2: Example of Hypothetical Data

Positive outlier:	Two study results, one with a mean 1 percentage point below the stated value and one with a mean 2 percentage points above the stated value; in the example, they would see the means 1 and 4.
Negative outlier:	Two study results, one with a mean 2 percentage points below the stated value and one with a mean 1 percentage point above the stated value; in the example, they would see the means 0 and 3.
No CIs:	No confidence intervals are provided.
Small CIs:	Confidence intervals are provided that extend 2 percentage points above/below each disaggregated data point.
Large CIs:	Confidence intervals are provided that extend 3 percentage points above/below each disaggregated data point.

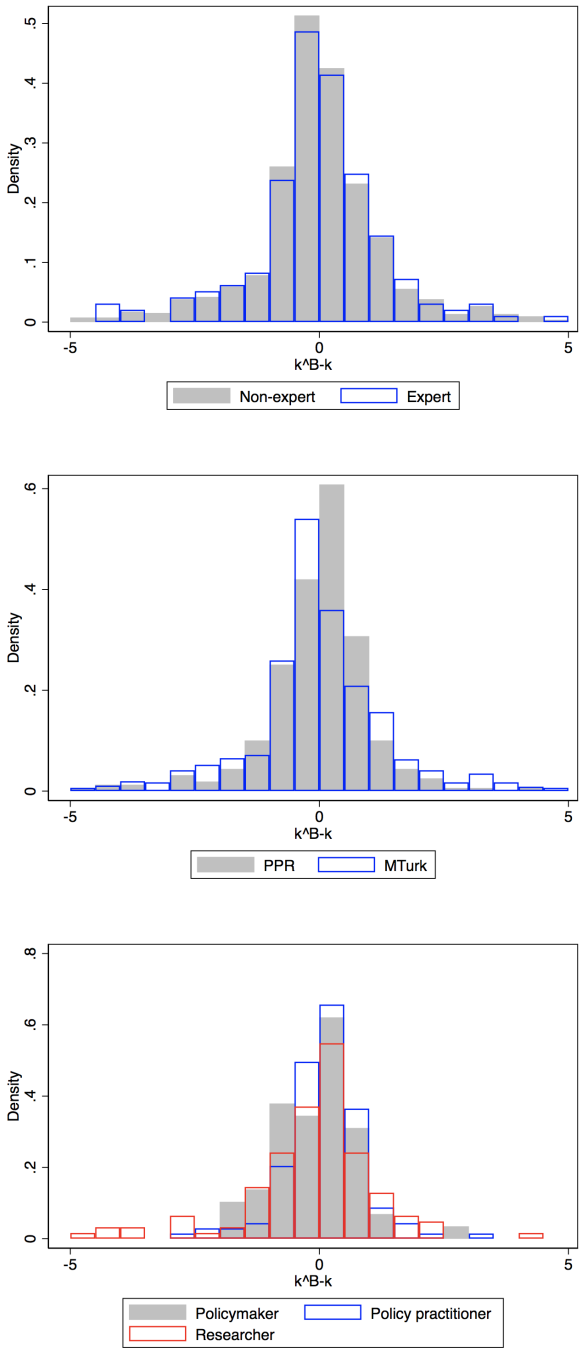
This description of the different types of evidence a participant could view is based on the hypothetical case of someone who previously reported they thought enrollment rates increased by 2 percentage points.

Figure A1: Distribution of Prior Means by Expertise



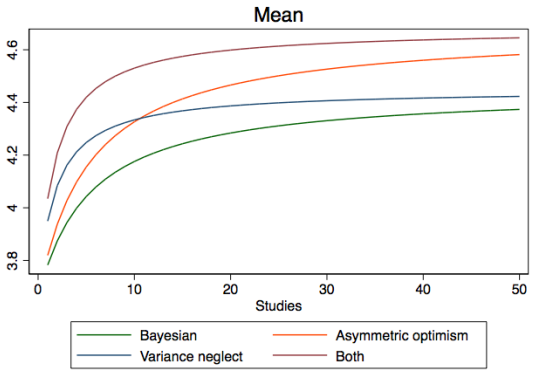
This figure plots values of prior means and standard deviations among those who had "never heard of it", "heard of it but never heard of any studies on it", or "heard of it and heard of some studies" (Non-expert) vs. those who had "heard of it and very familiar with studies" (Expert).

Figure A2: Distribution of $k^B - k$



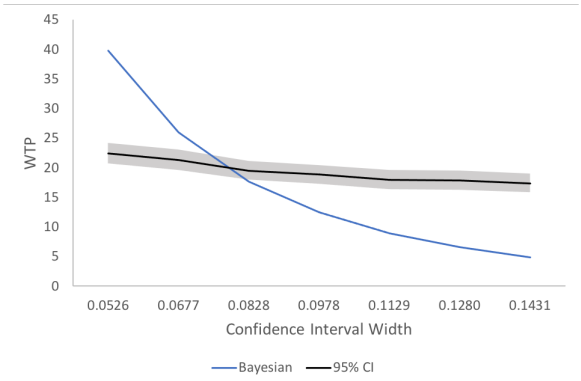
This figure plots values of $k^B - k$ calculated from respondents' reported μ_{t-1}, μ'_t, σ_{t-1} and the provided Y_i values. Values below -5 or above 5 are not included for legibility. The top plot distinguishes between responses to a "knowledge" question asked of all respondents: for each intervention, respondents were asked to specify whether they had "never heard of it", "heard of it but never heard of any studies on it", "heard of it and heard of some studies", or "heard of it and very familiar with studies". For visual clarity, the first three categories are collapsed into "Non-expert" and the last considered "Expert".

Figure A3: Simulated Evolution of Beliefs Over Time



This figure illustrates how individuals' posterior means would evolve over time, given the median estimated values of γ and λ for each subgroup, and assuming a distribution of study results as in the data in Figure 4. In particular, we separate out sampling variance from true inter-study variation in real study results using meta-analysis, and then we use the residual inter-study variance along with the mean to simulate a set of normally-distributed results. We assume each study has the typical standard error present in the real data. We then simulate drawing up to 50 new "study results" (treatment coefficients and standard errors). In generating each subgroup's results, we start from the same priors of μ_{t-1}=3.656 and v^2_{t-1}=1.413. In our parameterization, Bayesians and those suffering from variance neglect do not converge in beliefs even as the number of studies grows beyond the typical total number of studies for a topic in development economics. The beliefs of those suffering from asymmetric optimism never converge to those of a Bayesian, and those suffering from both asymmetric optimism and variance neglect have posterior expectations that diverge even more from a Bayesian's. When the number of studies is particularly small, as is often the case in development economics, the differences are relatively large.

Figure A4: Respondents' Willingness-to-Pay for Results vs. a Bayesian



This figure shows how the MTurk sample was insensitive to results with different 95% confidence intervals, compared to a Bayesian. The shaded area around the solid line indicates the 95% confidence interval. Prices are in terms of number of tokens. Notably, the curve looks strikingly like the theoretical prediction in Figure 1.

Table A3: Distribution of Prior Mean

	PPR		MTurk	
Prior mean	Frequency	Cumulative percent	Frequency	Cumulative percent
-5	0	0	0	0
-4	0	0	0	0
-3	0	0	3	0.2
-2	0	0	3	0.4
-1	1	0.2	4	0.66
0	15	3.14	25	2.31
1	48	12.57	120	10.21
2	108	33.79	249	26.61
3	110	55.4	210	40.45
4	59	66.99	155	50.66
5	89	84.48	294	70.03
6	30	90.37	134	78.85
7	13	92.93	130	87.42
8	19	96.66	101	94.07
9	12	99.02	62	98.16
10	5	100	28	100
Total	509		1,518	

This table provides the distribution of prior means for the policymakers, policy practitioners, and researchers sample (PPR) as well as the MTurk sample, for those passing all the screenings and tests. Notably, there are few responses below 0; most responses fall between 0 and 5. There is also some evidence of rounding: one of the most popular prior means, for both the PPR sample and the MTurk sample, is 5, with large weights also being placed on 2 and 3.

Table A4: Predictors of Normally Distributed Priors and Posteriors

	Indicator of Whether Priors and Posteriors are Normal					
	(1)	(2)	(3)	(4)	(5)	(6)
Policymaker	-0.076					-0.035
	(0.05)					(0.05)
Pol. Pract		0.010				0.002
		(0.03)				(0.04)
Expert			0.026			0.006
			(0.03)			(0.03)
Prior Mean				-0.045***		-0.044***
				(0.01)		(0.01)
Prior Variance					-0.008	-0.008
					(0.01)	(0.01)
Observations	602	602	602	602	601	601

This table reports the results of regressions of a binary variable capturing whether a respondent provided normally-distributed priors and posteriors to a given question on attributes of the individual and their responses, leveraging the policymakers, policy practitioners and researchers sample. "Expert" is a binary variable capturing whether, for a given intervention, respondents self-reported themselves as having "heard of it and very familiar with studies".

Table A5: Tests of Asymmetric Optimism: Heterogeneity by Expertise

	Normally-distributed			Any distribution		
	$I(k > 0.5)$	k		$I(k > 0.5)$	k	
	(1)	(2)	(3)	(4)	(5)	(6)
Good News	2.008***	0.075	0.178***	2.094***	0.086	0.169**
	(0.44)	(0.06)	(0.07)	(0.48)	(0.07)	(0.07)
Expert	0.937	-0.011	0.004	0.958	-0.022	0.006
	(0.27)	(0.07)	(0.07)	(0.28)	(0.07)	(0.08)
Good News *	1.090	-0.129	-0.018	1.026	-0.071	0.016
Expert	(0.48)	(0.13)	(0.14)	(0.48)	(0.14)	(0.15)
Observations	475	254	317	429	229	287
k restrictions	-	0-1	-0.5-1.5	-	0-1	-0.5-1.5

This table reports the results of regressions of k on an indicator of whether the respondent received "good news", *i.e.* positive new data relative to their priors. Columns (1) - (3) report results using the sample of policymakers, policy practitioners and researchers with normally-distributed priors and posteriors; Columns (4) - (6) report results using the sample of policymakers, policy practitioners and researchers regardless of their distributions of priors and posteriors. Columns (1) and (4) present the results of a logistic regression of $I(k > 0.5)$ on whether the respondent received good news, using exponentiated coefficients; Columns (2) and (5) regress k on whether the respondent received good news, restricting attention to only those observations for which $0 \leqslant k \leqslant 1$; Columns (3) and (6) also regress k on whether the respondent received good news, considering only those observations for which $-0.5 \leqslant k \leqslant 1.5$. Only those who provided prior means between 0-5 percentage points were included in the tests for asymmetric optimism, as we were unable to show new data above or below higher or lower priors without going out of range of what could be displayed. Including those outside of this range could introduce bias. Nonetheless, results are comparable when those outside this range are included, and those results are available upon request. "Expert" is a binary variable capturing whether, for a given intervention, respondents self-reported themselves as having "heard of it and very familiar with studies".

Table A6: Tests of Variance Neglect: Heterogeneity by Expertise

	$I(k^B - k > 0)$	$k^B - k$	
	(1)	(2)	(3)
Large C.I.	1.200	-0.156**	-0.136*
	(0.29)	(0.06)	(0.07)
Expert	0.755	-0.083	-0.148
	(0.26)	(0.10)	(0.11)
Large C.I. *	1.677	0.218	0.309*
Expert	(0.94)	(0.15)	(0.16)
Observations	349	206	256
k restrictions	-	0-1	-0.5-1.5

This table reports the results of regressions of $k^B - k$ on an indicator of whether the respondent saw large confidence intervals as opposed to small confidence intervals. Respondents can be included regardless of their priors, unlike in testing for asymmetric optimism, but cases in which respondents were randomized into seeing no confidence intervals are excluded. Columns (1) - (3) report results using the policymakers, policy practitioners and researchers sample. Column (1) presents the results of a logistic regression of $I(k^B - k > 0)$ on whether the respondent saw a large confidence interval as opposed to a small confidence interval, using exponentiated coefficients; Column (2) regresses $k^B - k$ on whether the respondent received a large confidence interval as opposed to a small confidence interval, restricting attention to only those observations for which $0 \leqslant k \leqslant 1$; Column (3) also regresses $k^B - k$ on whether the respondent received a large confidence interval or a small confidence interval, considering only those observations for which $-0.5 \leqslant k \leqslant 1.5$. Again, cluster-robust standard errors are used. All results are on the sample of those who have normally distributed priors and posteriors, as k^B is calculated under that assumption. "Expert" is a binary variable capturing whether, for a given intervention, respondents self-reported themselves as having "heard of it and very familiar with studies".

B MTurk Sample Results

As described in the main text, we ran the main experiment on a supplemental MTurk sample to understand how general the biases we observed were. We required a HIT Approval Rate (%) for all Requesters' HITs greater than or equal to 95 and Number of HITs Approved greater than or equal to 50.[48] 1,600 responses were solicited. In contrast to the policymakers, policy practitioners and researchers, who were interviewed one-on-one, the MTurk workers took the survey unsupervised.

MTurk participants were offered $1.50 for the relatively long survey. We were concerned that without incentivizing thoughtful responses, participants might not put in the effort to understand and carefully answer the questions. However, not incentivizing the responses would provide greater comparability with the results from policymakers, policy practitioners and researchers. Thus, we continued to not incentivize responses but chose to implement pre-specified screening questions to filter out inattentive participants. These screening questions were described in a pre-analysis plan posted on the Open Science Framework (OSF) and the study pre-registered at the AEA RCT Registry.[49] The screening questions are described in more detail in the next section.

For the MTurk sample, if we could not calculate k for either of the two sets of questions for which we tried to calculate it, we dropped the response and recruited a new participant.[50]

As in the policymakers, practitioners and researchers sample in the main text, it may be helpful to consider how many priors and posteriors were normally distributed. In 3.6% of cases, respondents put weight in only one or two bins, and we can neither prove nor disprove that the priors are normally distributed. Kolmogorov-Smirnov tests reject an additional 14.5% as non-normal.[51] Among those individuals who had priors spanning at least three bins and which could not be rejected as normal by a Kolmogorov-Smirnov test, 2.8% had posterior distributions that fell in 1-2 bins and a further 7.1% of the posterior distributions were rejected as not normal by a Kolmogorov-Smirnov test.

Overall, 76% of the MTurk respondents reported prior and posterior distributions that were in three or more bins and consistent with being normally distributed according to a Kolmogorov-Smirnov test. Only 44% of the MTurk estimates for k fall between 0 and 1.

[48] A HIT is a single MTurk task, *e.g.* a past exercise that they completed.

[49] OSF repository: https://osf.io/2da9p/, AEA registration id: AEARCTR-0001237, https://www.socialscienceregistry.org/trials/1237.

[50] For the policymakers, policy practitioners and researchers sample, incomplete surveys were not discarded due to the relative difficulty of obtaining these responses.

[51] As in the PPR sample, k could not be calculated for a small share (8.1%) of observations for the mechanical reason that the point estimate that respondents were shown, which was based on the first mean value that they stated, turned out to be exactly equal to the mean that we calculated from their putting weight in bins.

B.1 Screening Criteria

As described, MTurk responses were subject to screening. In particular, the first question asked what respondents thought the likelihood was that it would rain tomorrow in their city. They were then asked: "Now suppose that the weather forecast says there is a 50% chance it will rain tomorrow. Now what do you think is the likelihood that it will rain tomorrow?" If their new estimate was outside of the range between their initial answer and 50%, they were excluded, with an exception that will be described below. This rule excluded those whose responses implied $k < 0$ (*e.g.* they initially answer 10%, then update their answer to 0%, or if they initially answer 90%, then update their answer to 100%) and those whose responses implied $k > 1$ (*e.g.* they initially believe the likelihood is 10%, then update their answer to 60%, or they initially believe the likelihood is 90%, then update their answer to 40%). The second screening question asked what respondents thought the average monthly temperature would be in Paris this month. Again, they were provided with new information and those who provided a second answer that implied $k < 0$ or $k > 1$ were excluded, barring the exception described below. The third screening question presented them with pre-populated sliders that put probability weights on low temperature ranges. They were asked to modify these weights given the new information that two women who were perfectly informed as to what the weather would be like decided to wear shorts and a T-shirt; presumably, this would imply a higher temperature than the low numbers provided. Anyone who modified the sliders so as to result in a still lower mean temperature was excluded, barring the exception described below. Finally, we also excluded anyone who failed to shift the pre-populated sliders in the third question to focus on those who put some effort into answering the questions.

The point of these questions was not to screen out people who suffer from gambler's fallacy or hot hand fallacy or who updated in some other way so as to result in implied values of $k < 0$ or $k > 1$. Rather, if an MTurk worker answered in this way with regards to something as familiar as the weather, it seems likely that they were simply not paying attention to the question. By repeating the same kind of screening question three times, we can detect whether someone is consistently answering in a way that would imply $k < 0$, consistently answering in a way that would imply $k > 1$, or answering inconsistently across questions. Hence, we allowed one exception to the above screening rules: if someone consistently updated in a way that would imply $k < 0$ through all screening questions or in a way that would imply $k > 1$ through the first and second question and not $k < 0$ in the third question (as we cannot detect if $k > 1$ in the third question), we included them in the sample. In total, of 1,675 MTurk respondents[52], 1,183 passed the requirement to not

[52]We accidentally gathered slightly more data than initially planned, as a few more people answered the survey than filled in a survey code on MTurk within the alloted time, such that the HIT was not counted

update very differently across several questions, *i.e.* to not answer as though $k > 1$ for some questions and as though $k < 0$ for other questions. Of these, 1,029 met the other screening criteria.

Both the policymakers, policy practitioners and researchers sample and the MTurk sample were asked at the end of the introductory section if they understood how to use and interpret the slider bars, and if anyone selected the response "No", they were excluded from the sample.

B.2 MTurk Results

The MTurk sample had very similar results to the policymakers, practitioners and researchers sample. Results for this sample are provided in Tables B1-B3. As in the main sample, MTurk workers exhibited asymmetric optimism (Table B1) and variance neglect (Table B2). The MTurk sample also updated more when more information was provided (Table B3).

We did not ask allocation questions of the MTurk sample, in the interest of time.

and was re-offered to other participants.

Table B1: Tests of Asymmetric Optimism: Regressing k, MTurk Sample

	Normally-distributed			Any distribution		
	$I(k > 0.5)$	k		$I(k > 0.5)$	k	
	(1)	(2)	(3)	(4)	(5)	(6)
Good News	1.766***	0.091***	0.118***	1.821***	0.089***	0.150***
	(0.23)	(0.03)	(0.04)	(0.22)	(0.03)	(0.04)
Observations	1053	416	594	1189	457	661
k restrictions	-	0-1	-0.5-1.5	-	0-1	-0.5-1.5

This table reports the results of regressions of k on an indicator of whether the respondent received "good news", *i.e.* positive new data relative to their priors. Columns (1) - (3) report results using the sample of MTurk workers with normally-distributed priors and posteriors; Columns (4) - (6) report results using the sample of MTurk workers regardless of their distributions of priors and posteriors. Columns (1) and (4) present the results of a logistic regression of $I(k > 0.5)$ on whether the respondent received good news, using exponentiated coefficients; Columns (2) and (5) regress k on whether the respondent received good news, restricting attention to only those observations for which $0 \leqslant k \leqslant 1$; Columns (3) and (6) also regress k on whether the respondent received good news, considering only those observations for which $-0.5 \leqslant k \leqslant 1.5$. Only those who provided prior means between 0-5 percentage points were included in the tests for asymmetric optimism, as we were unable to show new data above or below higher or lower priors without going out of range of what could be displayed. Including those outside of this range could introduce bias. Nonetheless, results are broadly comparable when those outside this range are included, and those results are available upon request.

Table B2: Tests of Variance Neglect: Regressing $k^B - k$, MTurk Sample

	$I(k^B - k > 0)$	$k^B - k$	
	(1)	(2)	(3)
Large C.I.	0.770**	-0.121***	-0.156***
	(0.10)	(0.03)	(0.04)
Observations	1025	454	639
k restrictions	-	0-1	-0.5-1.5

This table reports the results of regressions of $k^B - k$ on an indicator of whether the respondent saw large confidence intervals as opposed to small confidence intervals. Respondents can be included regardless of their priors, unlike in testing for asymmetric optimism, but cases in which respondents were randomized into seeing no confidence intervals are excluded. Columns (1) - (3) report results using the MTurk worker sample. Column (1) presents the results of a logistic regression of $I(k^B - k > 0)$ on whether the respondent saw a large confidence interval as opposed to a small confidence interval, using exponentiated coefficients; Column (2) regresses $k^B - k$ on whether the respondent received a large confidence interval as opposed to a small confidence interval, restricting attention to only those observations for which $0 \leqslant k \leqslant 1$; Column (3) also regresses $k^B - k$ on whether the respondent received a large confidence interval or a small confidence interval, considering only those observations for which $-0.5 \leqslant k \leqslant 1.5$. Again, cluster-robust standard errors are used. All results are on the sample of those who have normally distributed priors and posteriors, as k^B is calculated under that assumption.

Table B3: Impact of More Information, MTurk Sample

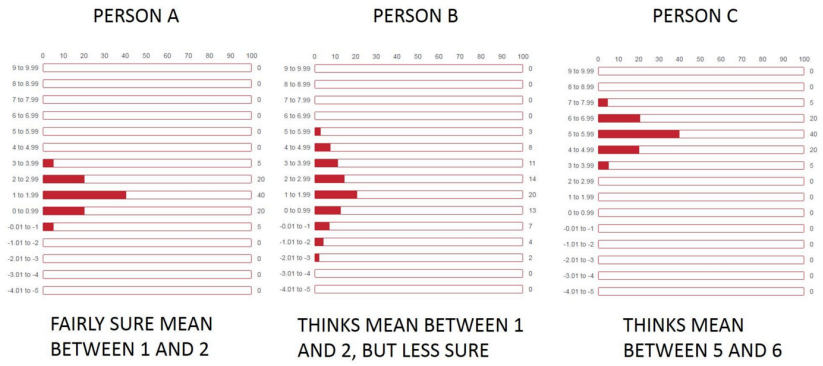
	$I(k > 0.5)$	k	
	(1)	(2)	(3)
Point Estimate	0.662**	-0.096***	-0.065***
	(0.12)	(0.02)	(0.03)
Point Estimate and	0.644**	-0.097***	-0.067***
Confidence Interval	(0.11)	(0.03)	(0.02)
Point Estimate,	0.978	0.010	-0.020
Confidence Interval, IQR	(0.18)	(0.02)	(0.02)
Observations	1985	1419	1830
k restrictions	-	0-1	-0.5-1.5

This table shows the impact of providing more information on k on the MTurk sample. Each type of information is represented by an indicator variable, with the category left out being shown point estimates, confidence intervals, the interquartile range, and maximum and minimum values. Column (1) shows the results of a logistic regression on $I(k > 0.5)$ with exponentiated coefficients. Column (2) restricts attention to $0 \leqslant k \leqslant 1$, and Column (3) restricts attention to $-0.5 \leqslant k \leqslant 1.5$.

C Experimental Details

Figure C1: Understanding Check

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



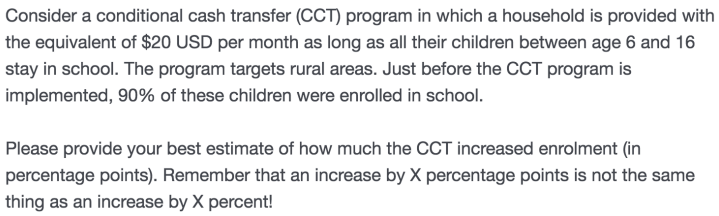
For instance, person A and B both suggest that the impact of a program is most likely to be in the range of 1 – 2 percentage points, while person C thinks the most likely range is between 3 – 4 percentage points.

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

Do these examples make sense to you?

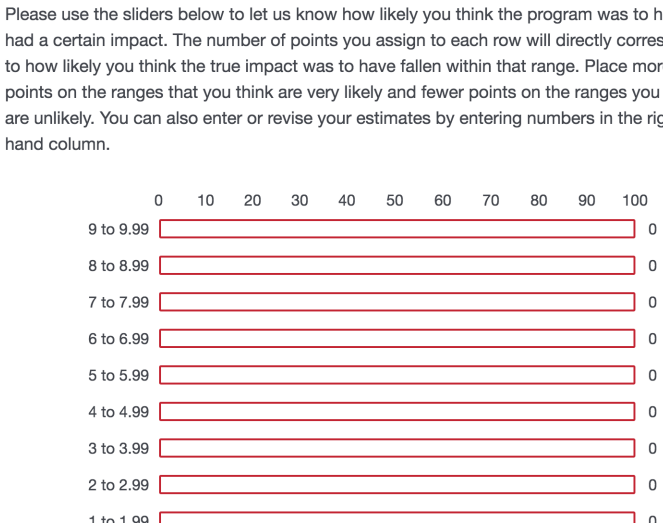
Respondents were walked through several examples of how they might distribute weights to different bins. MTurk respondents were provided with the accompanying written text describing each picture, while policymakers were provided with this information orally. We ran experiments showing different distributions in the video and understanding check (normal, uniform, and a mix of normal and uniform) to mitigate concerns that these distributions biased later responses. Results are available upon request.

Figure C2: Sample Program Description



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

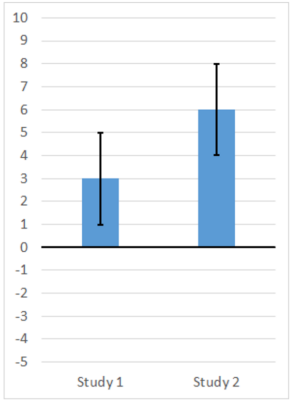
Figure C3: Assigning Likelihoods



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

Figure C4: Sample New Data

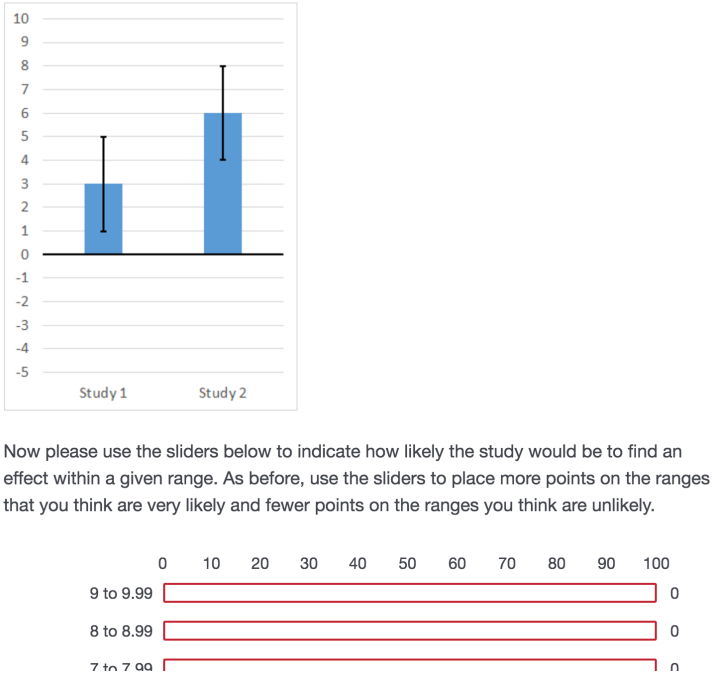
Suppose that 2 independent studies of this program were conducted. Each study followed the exact same design, but you do not know in which order they were done. One study found that the program increased enrolment by 3.0 percentage points, plus or minus 2.0 (this means that the 95% confidence interval was between 1.0 and 5.0 percentage points). The other study found that the program increased enrolment by 6.0 percentage points, plus or minus 2.0. A graphical depiction of the results of the 2 studies is provided below:



If a third study were done on this program following the exact same design, what effect do you think it would find? Please provide your best estimate.

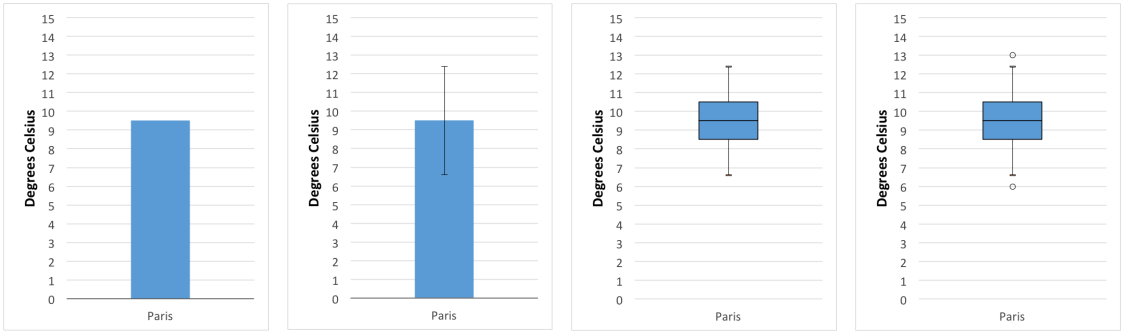
Respondents were then randomly shown data and asked to provide another estimate.

Figure C5: Assigning Likelihoods after Viewing New Data



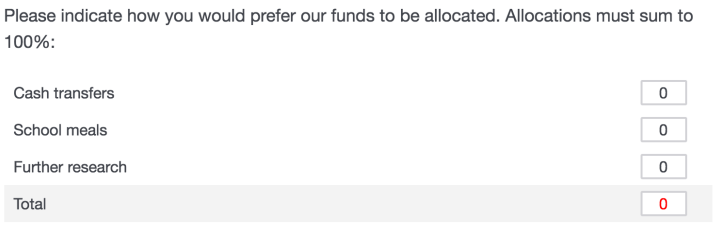
Respondents were also asked to provide their posteriors using slider bars.

Figure C6: Types of Information Provided for Information Experiment



Four types of information were provided in the information experiment in the introductory section of the survey: historical data was presented without confidence intervals, with confidence intervals, with confidence intervals and the interquartile range, and with confidence intervals, the interquartile range, and maximum and minimum values.

Figure C7: Allocation Question



Respondents were asked to allocate funds between three options: cash transfer programs, school meals programs, and further research.

Figure C8: Sample Screening Question

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