How Do Policymakers Update?

Eva Vivalt[*]

The Australian National University

Aidan Coville[†]

World Bank

Preliminary and Incomplete: Please Do Not Cite or Circulate

October 7, 2018

Abstract

Evidence-based policymaking requires not only evidence, but also for policymakers to update appropriately based on that evidence. We examine how policymakers, researchers, and others update in response to results from multiple studies, using a unique opportunity to run an experiment on policymakers. We find evidence of "variance neglect", a bias similar to extension neglect in which confidence intervals are ignored. We also find evidence of increased updating on results better than one's priors. Together, these results mean that policymakers might be biased towards those interventions with a greater dispersion of results. Finally, we test several possible ways to mitigate the observed biases.

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

[†] E-mail: acoville@worldbank.org.

1 Introduction

Evidence-based policymaking has the potential to improve many people's lives and many impact evaluations are motivated by the hope that they will inform policy decisions. Yet in order to make evidence-based decisions, three things must happen. First, there must be evidence on which policymakers can base their decisions. Second, they must accurately update their beliefs based on that evidence. Finally, they must have the willingness and capability to make their decisions according to what the evidence shows. Most research focuses on the crucial first step of generating new evidence and building the knowledge base. Some research also focuses on the last question; for example, considering the political economy issues affecting policy decisions (Persson and Tabellini, 2000; Krueger, 1993). In this paper, we focus on the second step.

It is well-known that there are a number of behavioural reasons why people may not accurately update their beliefs (Cameron, Loewenstein and Rabin, *eds.*, 2004; Kahneman, 2003). However, relatively few papers on behavioural biases have focused on policymakers. The closest work in this vein is Banuri *et al.* (2016), who survey World Bank staff members and staff at the Department for International Development (DFID) in the UK and find that they systematically mispredict how the poor will answer survey questions. A number of other papers also focus on prediction (DellaVigna and Pope, forthcoming). We focus specifically on how policymakers update in response to new information. We elicit their priors on the effects of a particular program, present them with new information from impact evaluations, and elicit their posteriors.

While there are relatively few papers focusing on policymakers, policymakers are an important population due to the potential weight of the decisions they make. In addition, we posit that some behavioural biases may be more relevant to them, for two reasons. First, they are subject to selection pressures that may result in their having slightly different biases than other populations. Second, they may be exposed to selected kinds of information that magnify these biases. In particular, policymakers are rarely provided with clear information about uncertainty.[1] If policymakers rarely face this information, they may have relatively less practice updating based on it.

We leverage a unique opportunity to run an experiment on policymakers, practitioners and researchers, in collaboration with the World Bank and the Inter-American Development Bank, that explores (i) whether updating biases exist and (ii) if so, whether the type of information provided may be able to reduce the impact of these biases. In particular, we hypothesize that policymakers are biased towards "good news", overweighting positive impact evaluation results compared to negative results (overconfidence). We also hypothesize that policymakers do not take the variance of impact evaluation results fully into consideration when updating (variance neglect).

We find evidence of both biases. We also test whether these biases can be mitigated by providing more information. Different treatment arms are provided with different amounts of information (*e.g.* various quantiles of the data). We find that providing more quantiles of the same distribution increases updating, suggesting that this can be used as a lever to alter how people perceive good or bad news and nudge them to more closely approximate Bayesian updating.

The "policymakers" on which we focus are not high-level policymakers. This is intentional; these people will rarely read academic papers thoroughly themselves and instead might rely on briefings from advisors. We instead focus on the policy-

[1] For example, the World Bank's flagship annual publication, the World Development Report, that is widely circulated among policymakers, was reviewed for the period 2010-2016. Of thousands of cited papers, information about standard errors or confidence intervals was only provided 8 times.

makers, practitioners, and researchers who attend World Bank and Inter-American Development Bank impact evaluation workshops; these attendees are particularly interested in impact evaluation and thus comprise an ideal sample for this study. The workshops are approximately one week long and designed as "matchmaking" events between those involved in development programs and researchers; government counterparts are paired with researchers and supposed to design a prospective impact evaluation for their program over the course of the week. Workshop participants include monitoring and evaluation specialists within government agencies; program officers in government agencies; World Bank or IDB operational staff; and other international organization operational staff such as technical advisors at USAID or DFID. The sample also includes a group of researchers, both from academic institutions as well as some international organizations such as the World Bank. In addition, we ran the experiment at the World Bank and Inter-American Development Bank headquarters in Washington, D.C., getting a broader swath of international organization staff. Finally, we separately ran the experiment on a sample of respondents through Mechanical Turk (MTurk) for an additional comparison group.

The rest of the paper proceeds as follows. First, we discuss the biases that we are studying and incorporate them in a quasi-Bayesian model of updating. We then describe the sample and the experiment. Results are then presented and discussed.

2 Model

2.1 Bayesian Updating

A policymaker might be 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 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 also has the opportunity to observe a signal about the effect of the program, Y_i with some normally distributed noise, $\varepsilon_i \sim N(0, \sigma_i^2)$. Y_i can be thought of as an impact evaluation result and can be written as:

$$Y_i = \theta_i + \varepsilon_i \tag{2}$$

Y_i therefore has variance $\tau^2 + \sigma_i^2$, which we will write as v_i^2. In the meta-analysis literature, this is known as a random-effects model.

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

$$\mu_t = \mu_{t-1} + k(Y_i - \mu_{t-1}) \tag{3}$$

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:

$$v_t^2 = \frac{v_{t-1}^2 v_i^2}{v_{t-1}^2 + v_i^2} \tag{4}$$

Similar equations could be written for the fixed effect model, in which $\tau^2 = 0$, substituting the sampling variance σ_i^2 for v_i^2. This can be thought of as the appropriate model for when one is considering information from replications.

In our experiment, we predominantly focus on the fixed effect case, framing the new information as coming from replications. This is to avoid the estimation challenges posed by estimating $\tau^2 \neq 0$, namely, that when we move away from a fixed effect world and introduce heterogeneous treatment effects, different people could build different mental models of how results depend on study characteristics (*e.g.* a mixed model), and we would have no way of separately estimating the model they have in mind. In order to avoid this problem, even when we present information from different settings (*i.e.* that might be subject to heterogeneous treatment effects) we will not be able to provide study details that could be used to build a more refined model; all studies are either described as replications or are otherwise exchangeable.

These equations assume that both the prior and the new information are normally distributed. We will be able to observe whether respondents have normally distributed priors and we will also observe the posteriors; we will focus on those cases in which both the prior and posterior are normally distributed, on the assumption that if both the prior and posterior are normally distributed, it is likely the respondents believed the data to be normally distributed as well.

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

2.2 Biases

2.2.1 Overconfidence

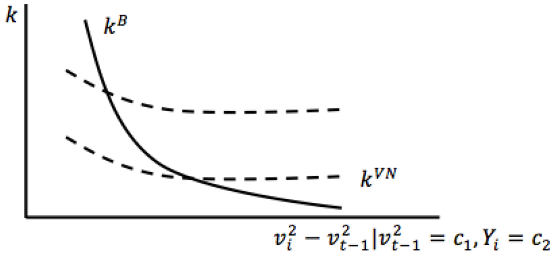
By overconfidence, we follow the literature and mean updating more on "good news" - information better than one's priors - than "bad news". We do not consider the bias of being overly certain, which is an alternative definition of overconfidence. Several different kinds of overconfidence with respect to good news may exist. One hypothesis is that people are overconfident with respect to point estimates, *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 $Y_i = 5$. It is also possible that, seeing a point estimate with a confidence interval, someone could believe that the true effect was closer to the upper bound rather than the lower bound of the confidence interval. In this example, the person would be overconfident with respect to confidence intervals. Other variations may also exist.

We distinguish between these types of overconfidence because they have practical significance. Under the second type of overconfidence, providing confidence intervals could actually worsen updating relative to not providing confidence intervals for a wide range of confidence interval values. For example, if someone is overconfident with respect to confidence intervals and views a confidence interval of 0 to 10, they might update as though they perceived a signal that it was 10.

Policymakers might be particularly prone to overconfidence because of the nature of their work: constantly having to sell or promote their programs. There is also often not much of a culture of dissent in large, bureaucratic organizations. It is also possible that overconfidence is actually selected for in the process of becoming a policymaker, development practitioner, or researcher.

To model overconfidence formally, we adapt Rabin and Schrag's classic paper on confirmation bias (1999), 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 some of the time they misperceive one signal as the other one. To expand this to a world in which many signals could be perceived and we are interested in overconfidence, we might say that people observing a signal Y_i perceived that they saw $Y_i + \gamma$ for some $\gamma > 0$. This would have the effect of resulting in a calculated k greater

Figure 1: Baysian 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.

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.

2.2.2 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 with:

$$k = \frac{v_{t-1}^2}{v_{t-1}^2 + (v_i^2 + \lambda)} \tag{5}$$

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

They could also suffer an error in updating the variance itself:

$$v_t^2 = \frac{v_{t-1}^2(v_i^2 + \lambda')}{v_{t-1}^2 + (v_i^2 + \lambda')} \tag{6}$$

where λ' is not necessarily equal to λ. Again, in the fixed effect case, v^2 is simply the sampling variance, σ^2.

For simplicity, we will focus on how policymakers form k. Later, we can consider how they update the variance.

Variance neglect is closely related to sample size neglect or, more broadly, extension neglect, but it is not quite the same thing. 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.

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 seeming variance neglect, however, there are also other potential causes of variance neglect, and later we will see that prospect theory alone is inconsistent with some of our results.

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 data repeatedly but a few data points once.

The next sections provide more information on the data and methods that will be used.

3 Sample

This study leverages a unique opportunity to run an experiment on policymakers, practitioners, and researchers in international development.

In particular, we interview individuals who previously attended World Bank or Inter-American Development Bank workshops in which policymakers, practitioners, and researchers were invited to learn more about impact evaluation and design an impact evaluation. Data were collected at 7 World Bank workshops run by the Development Impact Evaluation (DIME) research group. The workshops were conducted in Mexico City (May 2016, March 2017), Nairobi (June 2016), Lagos (May 2017), Washington, DC (May 2017, June 2017), and Lisbon (July 2017). The first two workshops were used as pilots to refine the questions and prior/posterior elicitation mechanisms. Data were also elicited at 2 Inter-American Development Bank (IDB) workshops in Washington, DC in June, 2017, and May, 2018.

These conferences attract participants from around the globe, rather than just participants from the host country or region. 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.

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

Individuals were surveyed by enumerators during breaks in the workshops. Of 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. During the pilots, individuals were allowed to take the survey by themselves on tablets we provided and, given that many could take the survey at the same time, we had a 95% response rate. However, we changed approaches after the pilot in favor of one-on-one enumeration to reduce noise due to participants' lack of familiarity with operating the tablets and to increase attentiveness. After making this change, we still had overwhelming interest in the survey among attendees but, being limited to the breaks in the workshops, only managed to survey an average of 23 per workshop. Breaks were roughly the duration of the survey, and lunch might span 2-3 times the length of the typical break, depending on workshop timing. Thus, this response rate represents essentially the maximum number of responses that could be gathered in the allotted time. We may expect that those who managed to take the survey may have been particularly interested in taking it or quick to approach the enumerators during a break, but we have no reason to believe that this represents a substantially different population than the universe of conference attendees. Response rates are detailed by workshop in Table 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 912 eligible past workshop attendees, 66 (7%) participated in the survey. Finally, participants were also recruited at the World Bank's headquarters and at the IDB's headquarters in Washington, D.C. A table was set up by the cafeteria and passers-by were able to take the survey with a trained enumerator. 125 World Bank responses and 23 IDB responses were collected in this manner over 24 days or 10 lunches, respectively;[2] enumerators covered lunch at the IDB but full or half-days at the World Bank. Summary statistics about the various recruitment strategies and the breakdown of participants by category (policymaker, practitioner, researcher) are provided in Table 2.

The experiment was set up to elicit two sets of priors and posteriors, so that with 376 respondents we would expect 752 priors and 752 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 intervention, resulting in 704 priors and posteriors from 352 respondents, or 94% of the total possible responses.

Finally, a set of responses was elicited on Mechanical Turk to provide a compar-

[2] Excluding 3 responses from support staff at the World Bank and 2 responses from support staff at the IDB. These did not meet our inclusion criteria but we could not bar them from participating upfront in this context.

Table 1: Participants at Workshops

	Eligible Attendees	Surveyed	Response Rate
Mexico, May 2016 (pilot)	107	105	0.98
Kenya, June 2016 (pilot)	48	43	0.90
Mexico, March 2017	93	34	0.37
Nigeria, May 2017	75	39	0.52
Washington, DC, May 2017	44	15	0.34
Washington, DC, June 2017 (IDB)	62	10	0.16
Washington, DC, June 2017	76	19	0.25
Portugal, July 2017	125	31	0.25
Washington, DC, May 2018 (IDB)	51	14	0.27
Total	526	162	0.31

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

Table 2: Respondents by Recruitment Strategy

	Policymakers	Practitioners	Researchers	Total
Workshops	0.38	0.31	0.30	162
Post-workshop videoconferences	0.17	0.29	0.55	66
Headquarters surveys	0.04	0.57	0.39	148
Total	0.21	0.41	0.38	376

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

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

Incentives for each of these groups are described in the experimental design section.

4 Experimental Design

4.1 Overconfidence and Variance Neglect

The overall structure of the experiment is simple: elicit participants' priors, randomly show them new information, and then elicit posteriors. This section describes the experimental design in more detail.

To elicit priors, participants were shown a short description of an intervention (example provided in Figure A.4 in the Appendix) and asked what they thought the effect of the program was on enrolment rates in percentage points. We then asked them to put weights on different possible effects the intervention could have had by having the enumerator drag slider bars next to different ranges (Figure A.5). We took this distribution as their prior. Before using these sliders, participants were shown a video describing how to use the sliders and were walked through an example about predicting the weather in order to be sure that they understood the exercise. At the end of this introduction, participants were asked if they understood and were only allowed to participate further if they stated that they did (Figure A.3). Only one participant stated that they did not understand the instructions and was prohibited from continuing.

The introductory text to the main experimental section suggested that enrolment rates were currently at 90 percentage points, so the most that respondents could reasonably expect the intervention to improve enrolment rates was an additional 10 percentage points; we also allowed them to guess negative values down to -5.

After participants provided their priors, they were randomized into seeing one of several sets of "new data", being asked to imagine that these data represented replications of studies on the same program.

In particular, respondents were presented with "new data" based on two hypothetical study results. These two studies either featured a positive or negative outlier relative to their stated prior, and either no confidence intervals, small confidence intervals or large confidence intervals. While the positive or negative outlier was randomly selected but influenced by their stated prior (such that a positive outlier was always a constant amount above the prior mean, and vice versa for a negative outlier), the confidence intervals that were provided bore no relation to the prior.

These questions are described in more detail in Table 3, using the example of someone who previously reported they thought enrolment rates increased by 2 percentage points.

Since these different possibilities could result in confidence intervals stretching up to 5 above or below the initial value that they provided, we could only follow the above strategy for those who initially state 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 and most responses indeed fell within 0 to 5 percentage points, especially given that respondents knew that baseline enrolment rates were 90 percentage points. Table 4 shows the distribution of prior means. To deal with any other responses, those who stated expected values

Table 3: Illustrative Example of Hypothetical Data

Positive outlier:	Two data points, 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 data points, 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. For the meta-analysis results, these are aggregated as they would be in a fixed effects meta-analysis.
Large CIs:	Confidence intervals are provided that extend 3 percentage points above/below each disaggregated data point. For the meta-analysis results, these are aggregated as they would be in a fixed effects meta-analysis.

The description above is based on the hypothetical case of someone who previously reported they thought enrolment rates increased by 2 percentage points.

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 overconfidence, since *e.g.* people who stated expected values greater than 5 will tend to see new data lower than their priors. Since the data are not symmetric, if we saw them updating less on these values, we would be unable to attribute that 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 robustness, for tests of overconfidence we will restrict attention to those whose priors fell between 0 and 5. On the other hand, we can still use all data for tests of variance neglect.

All data were presented as bar charts, and the order in which the two data points were provided (left to right) was randomized. The data were also described in the text. An example is provided in Figure A.6. 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_{t-1}^2 and μ_t in this manner, and given that we experimentally provide them with Y_i and σ_i^2, we can calculate what their k is using $\mu_t = \mu_{t-1} + k(Y_i - \mu_{t-1})$.

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 overconfidence. 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.

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 $\partial k^B / \partial \sigma_i^2 > \partial k / \partial \sigma_i^2$, where k^B represents the k of a Bayesian updater,

Table 4: Distribution of Prior Mean

Prior mean	PPR Frequency	 Cumulative percent	MTurk Frequency	 Cumulative percent
-5	0	0	0	0.0
-4	0	0	0	0.0
-3	0	0	3	0.2
-2	0	0	3	0.4
-1	1	0.2	4	0.7
0	15	3.4	25	2.3
1	46	13.0	120	10.2
2	101	34.1	249	26.6
3	106	56.3	210	40.5
4	58	68.4	155	50.7
5	82	85.6	294	70.0
6	27	91.2	134	78.9
7	11	93.5	130	87.4
8	17	97.1	101	94.1
9	10	99.2	62	98.2
10	4	100.0	28	100.0
Total	478		1,518	

This table provides the distribution of prior means for the policymakers, 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.

that would be evidence of variance neglect. We can then estimate λ.

One thing we cannot do is to both estimate overconfidence with respect to confidence intervals as well as estimating variance neglect, since confidence intervals derive from the variance. However, we may still be able to discern whether people are overconfident with respect to confidence intervals, if this outweighs variance neglect. For example, imagine that someone is presented with data whose mean is above their prior μ_{t-1}, with alternatively small or large confidence intervals. Someone who is overconfident with respect to confidence intervals, and for whom this outweighs any variance neglect, would put more weight on the result with larger confidence intervals. Someone for whom variance neglect outweighs overconfidence with respect to confidence intervals would put less weight on the result with larger confidence intervals, even though they would put relatively more weight on the result with larger confidence intervals compared to the result with smaller confidence intervals than someone who was Bayesian updating. In other words, if k_S represents the k that someone who saw small confidence intervals gives, k_L the k that someone who saw large confidence intervals would give, and the superscripts VN, O and B represent variance neglect, overconfidence with respect to confidence intervals, or Bayesian updating, respectively, we might expect $k_S^B - k_L^B > k_S^{VN} - k_L^{VN} > 0 > k_S^O - k_L^O$.

As the model depends on having normally distributed priors and data, in all analyses we restrict attention to those who report normally distributed priors, though we report all other priors descriptively. We cannot be certain that respondents believe the new information from replications is normally distributed, although we would argue that if their priors are normally distributed it would make sense for them to also believe the new information to be normally distributed. To further guard against misspecification, we will restrict attention to those respondents whose posteriors are also normally distributed.

4.2 Real-Life Decisions and Inter-Study Variation

The experiment described above uses hypothetical data. We would also like to use real data and tie the updates to real-life decisions. For this component of the study, we allow participants to decide how a small amount of external funds will be allocated. In particular, participants receive the following text:

> You will now be presented with information on two types of programs that both seek to increase the percent of children enrolled in school: cash transfers and school meals programs.
>
> We have a small amount of money that you can help to decide how to distribute. After being shown data on either cash transfers or school meals programs, you will be asked to decide how you would allocate this set of funds between cash transfers, school meals programs, or further research. For example, you could decide to distribute 33% to cash transfers, 33% to school meals programs, and 34% to further research. At the end of this study, we will randomly select one person's responses and distribute the money according to how that person decided it should be distributed.

Participants are then shown real-life data and asked to allocate funds between the different options. Some participants are also asked to provide their priors and posteriors before or after seeing the data and before determining their preferred allocations. The real-life data are taken from AidGrade's data set of impact evaluation results (2016). The point estimates and confidence intervals of the data provided randomly vary across participants, with the point estimates being either one result showing a 1 percentage point increase and one result showing a 4 percentage point increase or one result showing a 2 percentage point increase and one result showing a 5 percentage point increase, and with confidence intervals of either 2 or 5 percentage points. Respondents are also told that the school meals evidence shows an increase in enrolment rates by 3 percentage points, plus or minus 4 percentage points.

This approach allows us to see how much these differences affect real-world allocations and, in turn, estimate how much allocations could change if respondents were not biased or if they received information that helped them overcome their biases. There may be a wedge between updating and allocating which reduces the marginal value of new information.

The section with hypothetical data is more tailored to estimating overconfidence and variance neglect, in that we are able to provide information based on the priors that individuals stated (for example, randomly providing point estimates a certain amount above or below their priors). Since in this section we are working with real data, we do not have that flexibility. Further, due to the concern that merely by forcing respondents to think about their priors and posteriors, we are affecting how they would form allocations, we do not collect priors and posteriors in this section. This section precedes the section using hypothetical data, and at its conclusion respondents are told that they are moving on to a new section of the survey and will not need to use any information that was previously given.

Finally, as the point estimate provided for the alternative option of school meals is fixed at 3, this alternative option could be thought of as a safer bet than the more explicitly risky lottery of 1 and 4 or 2 and 5. If we were willing to make assumptions about 1) how respondents' biases affect how they view the data and 2) that there are no other relevant differences between how respondents see cash transfers and school meals programs, we could even back out some very crude estimates of risk preferences.

4.3 Information Treatments

If policymakers are biased, are there ways of presenting information so as to help them make better decisions?

The model makes some predictions as to how providing confidence intervals might affect estimates. If respondents do not suffer from variance neglect and are not overconfident, they will Bayesian update in the absence of any other biases, and in this case providing confidence intervals is helpful. If respondents are not overconfident but do suffer from variance neglect, confidence intervals are less helpful in that respondents somewhat ignore them. If respondents are overconfident and do not suffer from variance neglect, confidence intervals could help or hurt depending on whether respondents are overconfident with respect to point estimates or with respect to confidence intervals. If respondents are overconfident and suffer from variance neglect, confidence intervals could still help, but less so; confidence intervals alternatively could still hurt if respondents were overconfident with respect to confidence intervals. Table 5 summarizes.

The previously described experiment, in which results are presented with or without confidence intervals, can only tell us whether it is better to provide confidence intervals or not under assumptions about the true study sampling variance. To further examine how the information that is provided may affect estimates, we provide respondents with one of several types of information in a separate treatment. 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, providing more or less information. Figure A.2 illustrates.

Table 5: Predictions

	No variance neglect	Variance neglect
No overconfidence	Confidence intervals are helpful	Confidence intervals less helpful
Overconfidence	Confidence intervals are helpful / confidence intervals could hurt	Confidence intervals less helpful / confidence intervals could hurt

4.4 Incentives

Policymakers, practitioners and researchers were simply 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 drawn at random and awarded an additional prize: a MacBook. We did not further incentive responses because we were concerned that policymakers in particular would fear giving a "wrong" answer, so we did not want to increase the salience of the possibility of answering "incorrectly" by offering incentives for "correct" answers. The same incentives were offered to participants at the World Bank headquarters.

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

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

4.5 Screening

MTurk responses were also subject to screening. In particular, the first question asked what they 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 they move in the opposite direction to their initial answer, implying 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%) or if they adjust their answer in a way that would imply 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%), they were excluded, with an exception that will be described below. The second screening question asked what they think the average monthly temperature will 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. Finally, the third screening question presented them with pre-populated sliders that put probability weights on fairly 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. Anyone who modified the sliders so as to result in a lower mean temperature was excluded, barring the exception described below.

The point of these questions was not to screen out people who suffer from gambler's fallacy or hot hand fallacy. Rather, if an MTurk worker answered in this way with regards to something as familiar as the weather, we believe that most of these people would simply not be 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. If someone consistently updates 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, though we still consider their answers to be perverse enough that they are not be part of our preferred specification. To exclude those who put minimal effort into answering the questions, we also excluded anyone who failed to shift the pre-populated sliders in the third question.

Policymakers, practitioners and researchers were not faced with these questions and were not subject to these constraints. Both the policymakers, practitioners and researchers sample and the MTurk sample were also were asked one question at the end of the introductory section that explicitly asked if respondents understood how to use and interpret the slider bars, and if anyone selected the response “No”, they were excluded from the sample. The screening questions were pre-specified in a pre-analysis plan posted at the Open Science Framework.

5 Results

5.1 Descriptive Statistics: Priors

This section describes how respondents' priors were distributed, with reference to both the policymakers, practitioners and researchers sample (hereafter referred to as PPR) and the MTurk sample.

Recalling that our identification strategy for k assumes normally distributed priors, we test for normality using a Kolmogorov-Smirnov test. This test is typically conducted for continuous distributions, rather than for data that falls in discrete bins, as in our case. However, the Kolmogorov-Smirnov test can be applied to discrete data with modifications. The intuition is that we find the continuous normal distribution that would best fit the binned data we observe if it were binned using the same bins as in our discrete data. We then bin those continuous data using the same bins and test whether these two discrete distributions are significantly different. We discard those fairly rare cases in which respondents put weight in only one or two bins (7.2% of the PPR sample, 3.6% of the MTurk sample).

While this is a natural approach, it should 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. The Kolmogorov-Smirnov tests of the prior distributions reject 10.0% of the remaining observations in the PPR sample.

Of 1,675 MTurk respondents[3], 1,029 passed the "screening" questions. Again, one set of screening questions required respondents to not update very differently across several questions, *i.e.* to not answer as though k >1 for some questions and as though

[3] 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 and was re-offered to other participants.

k <0 for other questions; 1,183 passed this requirement. There was also one screening question which provided respondents with a set of pre-filled slider bars and asked them to adjust the slider bars to reflect their beliefs given new information; since participants were not required to adjust the bars at all but the question was designed such that respondents should adjust their estimates upwards, we dropped anyone who did not adjust any of the bars, considering them too inattentive or uninterested in putting effort into their responses. After discarding the observations in which weight was placed in only one or two bins, as previously described, the Kolmogorov-Smirnov tests then discard an additional 15% of observations.

Finally, k could not be calculated for a subset of observations (11.8% in the PPR sample, 8.1% for the MTurk sample) 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.[4] 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. 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. For the PPR sample, incomplete surveys were not discarded, due to the relative difficulty of obtaining these responses.

5.2 Descriptive Statistics: Posteriors

The reported posteriors were broadly similar to the priors in terms of how many could be considered normal. Of those observations in the PPR sample for which the

[4] 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.

associated prior passed all the tests, 1.4% were dropped for having a posterior distributed in 1-2 bins and a further 4.7% of the posterior distributions were rejected as normal by a Kolmogorov-Smirnov test. For the MTurk sample, these numbers were 2.8% and 7.3%, respectively.

Overall, 79% of policymakers, practitioners and researchers and 76% of MTurk respondents reported distributions that are consistent with having normally distributed priors and normally distributed posteriors using a Kolmogorov-Smirnov test.

5.3 Descriptive Statistics: Distribution of k

Restricting attention to the group that reported normally distributed priors and posteriors, we calculated k. Figure 2 plots its distribution, illustrating a relatively large range with clusters of estimates around 0 and 1. It should be recalled that k should generally fall between 0 and 1, with those who take the data's mean as their posterior mean having k=1 and those who stick with their initial mean having k=0, however, only 55% of the PPR estimates and 44% of the MTurk estimates fall within this range.

Notably, MTurk workers updated more on the data than policymakers, practitioners or researchers. This makes intuitive sense: they also reported less familiarity with the types of interventions discussed. Figure 2 shows that policymakers, who also stated they were less familiar with studies than practitioners or researchers, also had higher values of k. Somewhat surprisingly, there was a cluster of researchers who appeared to update too much based on the data.

The bottom plot distinguishes between 4 possible responses to a "knowledge" question asked of all respondents: for each intervention, respondents were asked to specify whether they had "never heard of it" ("No Knowledge"), "heard of it but

Table 6: Quantiles of k

		Percentile				
	Subgroup	10	25	50	75	90
PPR	All	-1.00	0.00	0.52	1.02	2.33
	Policymakers	-0.29	0.00	0.61	1.22	2.43
	Practitioners	-1.00	0.00	0.30	1.00	1.72
	Researchers	-1.00	0.00	0.68	1.20	3.30
	Familiar	-0.71	0.00	0.62	1.03	2.43
	Unfamiliar	-2.48	0.00	0.30	1.05	3.35
MTurk	All	-1.51	-0.01	0.68	1.23	3.16
	Familiar	-2.36	-0.14	0.62	1.15	3.08
	Unfamiliar	-1.26	0.00	0.72	1.26	3.25

This table shows quantiles of k by subgroup. "PPR" refers to the policymakers, practitioners and researchers sample. "Unfamiliar" refers to those respondents who noted that they had either never heard of conditional cash transfers or school meals programs or had heard of them but had never heard of any studies on them, while "Familiar" refers to those who noted that they had heard of them and heard of some studies on them or had heard of them and were very familiar with studies on them.

never heard of any studies on it" ("Never Heart of Studies"), "heard of it and heard of some studies" ("Familiar"), or "heard of it and very familiar with studies" ("Very Familiar"). As expected, those who reported greater familiarity with a type of intervention updated less in response to new information, though differences are not large.

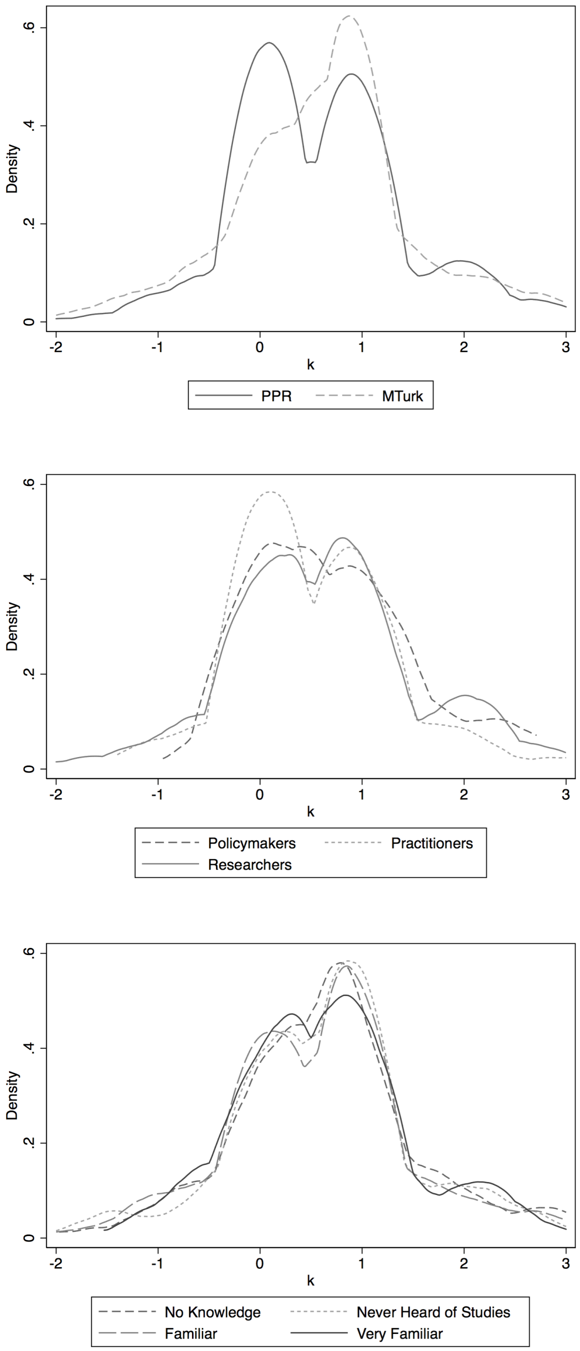
Our primary focus was on calculating k from the distributions that respondents provided. However, each time we asked for a prior or posterior, we also asked respondents to first provide their best guess of the effect of the program. We also use these as estimates of their μ_{t-1} and μ'_t in an alternative specification.

Table 6 shows quantiles of k for various types of respondent.

5.4 Tests for Biases

The wide dispersion of k values complicates testing for biases. In each of the PPR and MTurk sample, we restrict attention to alternative ranges of k: $0 \leqslant k \leqslant 1$,

Figure 2: 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 bottom plot distinguishes between 4 possible responses to a "knowledge" question asked of all respondents: for each intervention, respondents were asked to specify whether they had "never heard of it" ("No Knowledge"), "heard of it but never heard of any studies on it" ("Never Heart of Studies"), "heard of it and heard of some studies" ("Familiar"), or "heard of it and very familiar with studies" ("Very Familiar").

$-0.5 \leqslant k \leqslant 1.5$, and $-1 \leqslant k \leqslant 2$.

Table 7 presents results for regressions of respondents' seeming values of k on whether they received the "positive" news treatment. Robust standard errors are used, clustering observations at the individual level. Receiving the positive treatment significantly affected k for most specifications in the PPR sample and all specifications in the MTurk sample.

Table 7: Tests of Overconfidence

	PPR			MTurk		
	(1)	(2)	(3)	(4)	(5)	(6)
Positive	0.049	0.153**	0.300***	0.091***	0.118***	0.138**
	(0.06)	(0.06)	(0.08)	(0.03)	(0.04)	(0.06)
N	218	271	315	416	594	703
R^2	0.00	0.02	0.05	0.01	0.01	0.01

This table reports the results of regressions of k on an indicator of whether the respondent saw "positive" point estimates. "Positive" here means relative to their priors. Columns (1) - (3) report results using the policymakers, practitioners and researchers sample; Columns (4) - (6) report results using the MTurk sample. Columns (1) and (4) consider only those observations for which $0 \leqslant k \leqslant 1$; Columns (2) and (5) consider only those observations for which $-0.5 \leqslant k \leqslant 1.5$; Columns (3) and (6) consider only those observations for which $-1 \leqslant k \leqslant 2$. Only those who provided prior means between 0-5 percentage points were included in the tests for overconfidence, as we were unable to show new data above or below higher or lower priors without going out of range.

Recall that to test for variance neglect, we need to consider what someone who was Bayesian updating would do. Thus, in Table 8, 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 we showed them. The model implied that $k_S^B - k_L^B > k_S^{VN} - k_L^{VN}$; we can test this by regressing $k^B - k$ on whether the respondent saw large or small confidence intervals.[5] In Table 8, we observe that $k_S^B - k_S$ is indeed generally greater than $k_L^B - k_L$, as we predicted.

Table 8: Tests of Variance Neglect

	PPR			MTurk		
	(1)	(2)	(3)	(4)	(5)	(6)
Large C.I.	-0.099*	-0.078	0.027	-0.121***	-0.156***	-0.198***
	(0.06)	(0.06)	(0.08)	(0.03)	(0.04)	(0.05)
N	194	239	270	454	639	742
R^2	0.01	0.01	0.00	0.03	0.02	0.02

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 overconfidence, but cases in which respondents were randomized into seeing no confidence intervals are excluded. Columns (1) - (3) report results using the policymakers, practitioners and researchers sample; Columns (4) - (6) report results using the MTurk sample. Columns (1) and (4) consider only those observations for which $0 \leqslant k \leqslant 1$; Columns (2) and (5) consider only those observations for which $-0.5 \leqslant k \leqslant 1.5$; Columns (3) and (6) consider only those observations for which $-1 \leqslant k \leqslant 2$.

5.5 Heterogeneity by Profession and Gender

We may also be interested in how results vary by sub-sample. We consider heterogeneity by profession in Table 9 and heterogeneity by gender in Table 10.

In Table 9, we observe that, depending on the specification, practitioners and researchers updated less on the data than MTurk workers. This would be consistent with a story in which they had more background knowledge or narrower priors, and it does not require a difference in updating. Policymakers, practitioners and researchers do not appear to experience significantly more or less overconfidence than MTurk

[5] Again, cluster-robust standard errors are used.

workers, the sub-group left out. We also observe that policymakers, practitioners and researchers were no closer to Bayesian than MTurk workers. In almost all cases, they do not appear to suffer significantly more or less variance neglect than the MTurk workers. The one exception is that researchers appear to pay more attention to the variance.

In Table 10, we observe that in most specifications, women updated less based on the new information. We caution against over-interpreting this fact, because it could be an artefact of other features of the data. For example, men are over-represented among policymakers in our data set, contributing 66% of forecasts, while women are over-represented among researchers, contributing 54% of forecasts. For some specifications and samples, women appeared to suffer from more overconfidence and, in fact, be driving the results. For variance neglect, the only significant interactions showed the policymakers, practitioners and researchers sample exhibiting less variance neglect than the MTurk sample, with no differences apparent by gender.

5.6 Estimating γ and λ

We can also obtain distributions of γ and λ across individuals and test whether these distributions are different from 0. We find an average value of γ of 0.36 and an average value of λ of -2.23 in the PPR data, winsorizing the highest and lowest 1% of the data. γ is significantly different from 0 at $p < 0.05$ and λ is significantly different from 0 at $p < 0.001$.[6] If γ is positive, it indicates overconfidence; if λ is negative, it indicates that subjects update too much in response to the new information, given its confidence interval. In the MTurk data, the equivalent values for γ and λ are 0.54 and -1.78, significantly different from 0 at $p < 0.0001$ and $p < 0.01$, respectively.

[6] Without winsorizing, the values are still significant but likely noisier: the average value of γ is then 0.45, the average value of λ, -3.72, and they are significantly different from 0 at $p < 0.05$ and $p < 0.1$, respectively.

Table 9: Heterogeneity by Profession

	Overconfidence			Variance Neglect		
	(1)	(2)	(3)	(4)	(5)	(6)
Policymaker	-0.007	0.082	0.135	0.058	0.016	-0.147
	(0.09)	(0.09)	(0.12)	(0.09)	(0.11)	(0.12)
Practitioner	-0.158***	-0.148***	-0.169***	0.101	0.063	0.036
	(0.05)	(0.06)	(0.06)	(0.07)	(0.08)	(0.09)
Researcher	-0.094	-0.121*	-0.098	0.070	-0.012	-0.148
	(0.06)	(0.06)	(0.07)	(0.08)	(0.08)	(0.12)
Positive	0.091***	0.118***	0.138**			
	(0.03)	(0.04)	(0.06)			
Policymaker *	-0.145	-0.050	-0.019			
Positive	(0.15)	(0.16)	(0.18)			
Practitioner *	-0.032	0.004	0.154			
Positive	(0.09)	(0.10)	(0.14)			
Researcher *	-0.030	0.089	0.227			
Positive	(0.11)	(0.11)	(0.14)			
Large C.I.				-0.121***	-0.156***	-0.198***
				(0.03)	(0.04)	(0.05)
Policymaker *				0.016	0.009	0.214
Large C.I.				(0.14)	(0.14)	(0.14)
Practitioner *				0.024	0.105	0.121
Large C.I.				(0.09)	(0.10)	(0.12)
Researcher *				0.010	0.064	0.340**
Large C.I.				(0.11)	(0.11)	(0.16)
N	634	865	1018	648	878	1012
R^2	0.04	0.03	0.03	0.04	0.03	0.02

Columns (1) - (3) report the results of regressions of k on an indicator of whether the respondent saw positive new data relative to their priors. Columns (4) - (6) report 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; cases in which respondents did not view any confidence interval are excluded. Columns (1) and (4) consider only those observations for which $0 \leqslant k \leqslant 1$; Columns (2) and (5) consider only those observations for which $-0.5 \leqslant k \leqslant 1.5$; Columns (3) and (6) consider only those observations for which $-1 \leqslant k \leqslant 2$. The tests for overconfidence and variance neglect were conducted on slightly different samples. The tests of overconfidence 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).

Table 10: Heterogeneity by Gender

	Overconfidence			Variance Neglect		
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.048	-0.087*	-0.113*	0.017	0.037	0.032
	(0.04)	(0.05)	(0.06)	(0.05)	(0.06)	(0.07)
PPR	-0.069	-0.112*	-0.091	0.082	0.062	-0.066
	(0.05)	(0.06)	(0.07)	(0.07)	(0.08)	(0.10)
PPR * Female	-0.091	0.024	0.000	0.001	-0.075	-0.011
	(0.08)	(0.09)	(0.10)	(0.10)	(0.11)	(0.14)
Positive	0.058	0.033	0.097			
	(0.05)	(0.06)	(0.08)			
Female * Positive	0.075	0.190**	0.089			
	(0.07)	(0.09)	(0.11)			
PPR * Positive	-0.175*	-0.039	0.109			
	(0.09)	(0.10)	(0.13)			
PPR * Female *	0.315**	0.195	0.122			
Positive	(0.14)	(0.15)	(0.20)			
Large C.I.				-0.125***	-0.157***	-0.211***
				(0.05)	(0.06)	(0.07)
Female * Large				0.013	0.012	0.038
C.I.				(0.07)	(0.08)	(0.10)
PPR * Large				0.103	0.170*	0.312***
C.I.				(0.09)	(0.10)	(0.12)
PPR * Female *				-0.194	-0.206	-0.189
Large C.I.				(0.14)	(0.15)	(0.18)
N	634	865	1018	648	878	1012
R^2	0.06	0.04	0.03	0.04	0.03	0.02

"PPR" is an indicator of whether the respondent was in the policymakers, practitioners and researchers sample. Columns (1) - (3) report the results of regressions of k on an indicator of whether the respondent saw positive new data relative to their priors. Columns (4) - (6) report 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; cases in which respondents did not view any confidence interval are excluded. Columns (1) and (4) consider only those observations for which $0 \leqslant k \leqslant 1$; Columns (2) and (5) consider only those observations for which $-0.5 \leqslant k \leqslant 1.5$; Columns (3) and (6) consider only those observations for which $-1 \leqslant k \leqslant 2$. The tests for overconfidence and variance neglect were conducted on slightly different samples. The tests of overconfidence 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).

5.7 Changes in Allocations

In this section, we present reduced-form estimates of the impact of seeing more positive results data on allocations. Only policymakers, practitioners and researchers answered this part of the survey. In some contexts, we did not ask this part of the survey due to time constraints, so we only have allocations from 284 individuals.

There are several reasons we may expect participants' allocations to not be greatly affected by the information that they receive. First and foremost, in our experiment they only receive information about a particular outcome variable, and they may care about many different outcomes. Second, they may not feel like they have sufficient information to update much. As the section that asks participants to make allocations uses real rather than hypothetical data, no study details could be provided, lest they update on characteristics that vary 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 examines how a similar sample weights studies based on their characteristics, leveraging a discrete choice experiment.

The average allocations to CCT programs, school meals programs and further research, respectively, were 35%, 35% and 30%. Recall that in this part of the survey, respondents randomly viewed a selection of real data on CCT 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 CCTs of 8 percentage points; viewing results with larger confidence intervals resulted in a decrease in the amount allocated to CCTs of 8 percentage points. Interestingly, when respondents saw large confidence intervals, they allocated 5 percentage points more to further research. Results are presented in Table 11. Robust standard errors are used.

Table 11: Regressions of Allocations on Evidence Shown

	(1) Allocation to CCTs	(2) Allocation to CCTs	(3) Allocation to Research	(4) Allocation to Research
Large point estimate	7.625***		1.796	
	(1.99)		(2.51)	
Large confidence interval		-7.549***		5.078**
		(2.03)		(2.52)
Observations	284	284	284	284
R^2	0.05	0.05	0.00	0.01

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) and on 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). We were unable to ask the allocation question at all workshops due to time constraints, hence the smaller sample size.

It is interesting to consider how much allocations would change if respondents were Bayesian updating. We can approximate this by noting that, the way we parameterize it, overconfidence can be thought of as misperceiving a signal to think that the point estimate of the signal is higher than it really is, and variance neglect can be thought of as misperceiving a signal about the significance of new information. Recall that we estimated γ to be equal to 0.36 and λ to be equal to -2.23 in the PPR sample. If a policymaker suffering from overconfidence misperceived a signal that they thought a certain option had a point estimate that was 0.36 higher than a Bayesian might think it to be, the estimates in Table 11 would suggest they might allocate 2.7% more than a Bayesian would to that option. Similarly, suffering from variance neglect might allocate 11.0% more to that option than they would if they were Bayesian.

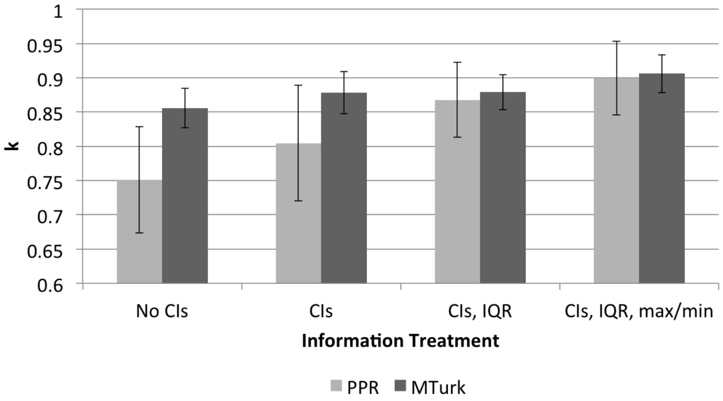
The obvious issue regarding this approach is that in equilibrium, this effect might be partially washed out. For example, if a policymaker wanted to allocate more to every program, they would not have enough money to do so, and so we might imagine they would still make the same allocations they would if they were Bayesian. However, there are two responses to this criticism. First, policymakers may be unlikely to receive information for many projects at the same time. For example, one could have priors about a default option and only observe a signal about one other program and then update too much based on those results. More fundamentally, it should be noted that due to the shape of the normal distribution, thinking that one perceives $Y_i + \gamma$ rather than Y_i results in more updating for larger values of Y_i, a fact that is obscured in the linear regressions of Table 9. The impact of λ on updating is also non-linear. Thus, while we present these back of the envelope calculations, it should be noted that they are necessarily particular to the details of the questions we asked and the priors respondents had.

5.8 How Much Information Should We Provide?

The type of information that was provided affected updating. Figure 3 shows results from the PPR and MTurk sample for each of the four treatment arms: providing results 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.

These k are winsorized at 5% to reduce noise. Both samples show some increased updating in response to more information. The PPR sample shows significantly

Figure 3: Updating by Type of Information Provided



This figure provides the values of k, and confidence intervals, for four types of information that were provided to the PPR and MTurk sample. These values of k were calculated from an introductory question on estimating the temperature.

greater updating when more information is provided; the differences are marginally insignificant for the MTurk sample but the point estimates follow a similar trend.

6 Conclusion

How policymakers update is an important topic, and this paper provides some first evidence using a unique opportunity to run an experiment with policymakers, practitioners and researchers. Respondents recruited through MTurk served as an additional comparison group.

We found that many people had seemingly normally distributed priors and posteriors, but the main parameter governing updating, k, often fell outside the standard range of 0 to 1. Few differences in updating were observed between policymakers, practitioners, researchers and MTurk workers, though policymakers, practitioners and researchers had narrower priors and consequently updated less in response to new information.

A model of quasi-Bayesian updating was built to accommodate overconfidence and variance neglect, and its parameters were estimated. These parameters were shown to be statistically significantly different from 0, the value they would take if respondents were fully Bayesian. We also found that new information affected allocation of resources, suggesting that if policymakers were Bayesian updaters they might allocate resources differently.

Finally, we saw that the amount of information provided matters. More information generally leads to increased updating. This suggests that in cases in which one has to share bad news, providing more detailed information may help.

References

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

Cameron, Colin, George Loewenstein and Matthew Rabin, *eds.* (2004). Advances in Behavioral Economics. New York: Russell Sage Foundation.

Banuri, Sheheryar, Stefan Dercon and Varun Gauri (2016). “The Biases of Policymakers”, working paper.

Banuri, Sheheryar, Stefan Dercon and Varun Gauri (2015). “The Biases of Development Professionals”, World Development Report, World Bank.

DellaVigna, Stefano and Devin Pope (forthcoming). “Predicting Experimental Results: Who Knows What?”, Journal of Political Economics.

DellaVigna, Stefano and Devin Pope (2018). “What Motivates Effort? Evidence and Expert Forecasts”, Review of Economic Studies, April 2018, Vol. 85, 10291069.

Eil, David and Justin Rao (2011). “The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself”, American Economic Journal: Microeconomics, vol. 3(2).

Kahneman, Daniel (2003). “Maps of Bounded Rationality: Psychology for Behavioral Economics”, American Economic Review, vol. 93(5).

Kahneman, Daniel and Amos Tversky (1979). “Prospect Theory: An Analysis of Decision under Risk”, Econometrica, vol. 47(2), pp. 263-291.

Krueger, Anne (1993). Political Economy of Policy Reform in Developing Countries. Cambridge: The MIT Press.

Ortoleva, Pietro and Erik Snowberg (2015). “Overconfidence in Political Behavior”, American Economic Review, vol. 105(2).

Persson, Torsten and Guido Tabellini (2000). Political Economics: Explaining Economic Policy. Cambridge: The MIT Press.

Rabin, Matthew and Joel Schrag (1999). “First Impressions Matter: A Model of Confirmatory Bias”, Quarterly Journal of Economics, vol. 114(1).

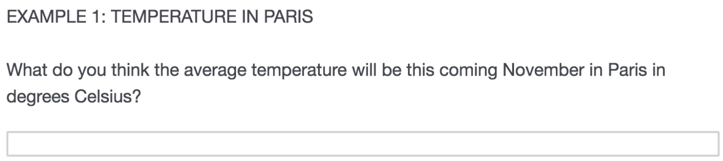
Rabin, Matthew and Dimitri Vayanos (2010). “The Gambler’s and Hot-Hand Fallacies: Theory and Applications”, Review of Economic Studies, vol. 77(2).

Appendix

Experimental Details

The following diagrams are excerpted from the survey.

Figure A1: Sample Screening Question

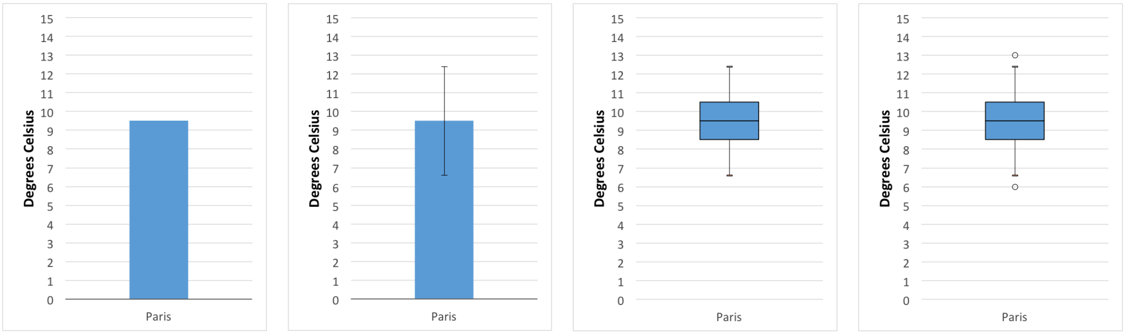


EXAMPLE 1: TEMPERATURE IN PARIS

What do you think the average temperature will be this coming November in Paris in degrees Celsius?

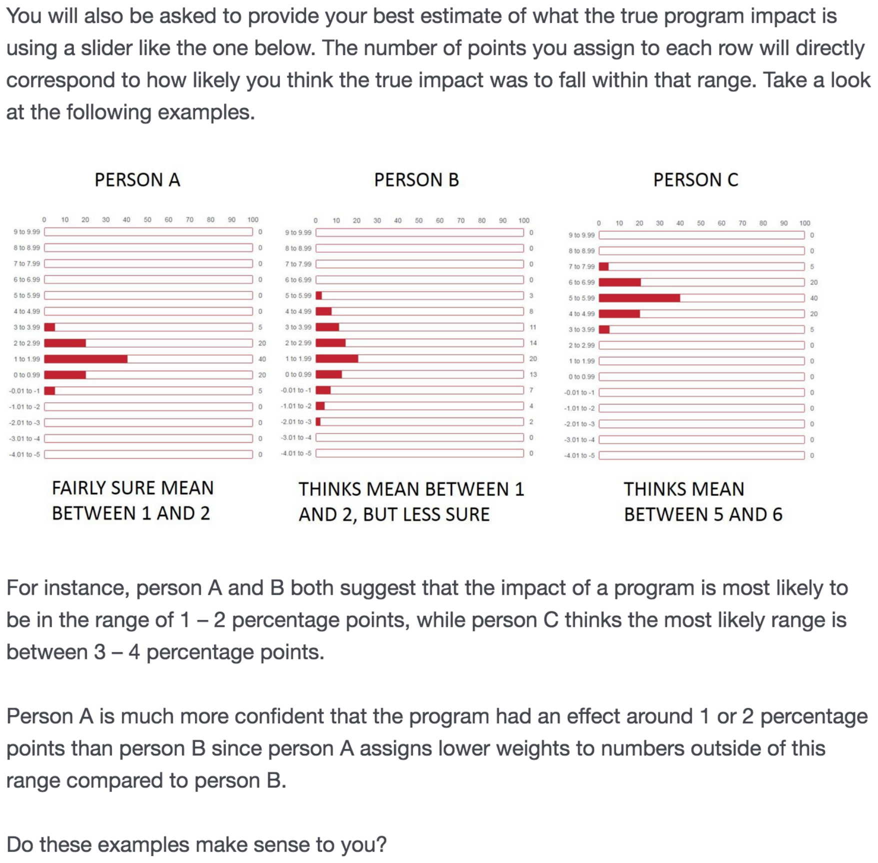
Several simple screening questions were used. After this question, respondents were presented with data and then asked to provide another estimate.

Figure A2: 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 A3: Understanding Check



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

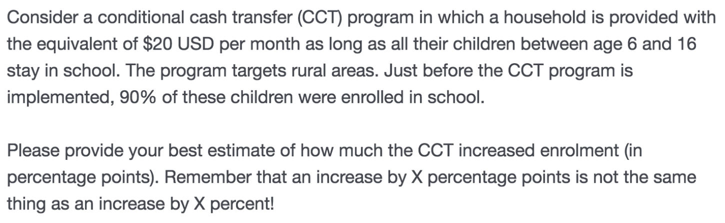
For instance, person A and B both suggest that the impact of a program is most likely to be in the range of 1 – 2 percentage points, while person C thinks the most likely range is between 3 – 4 percentage points.

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

Do these examples make sense to you?

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

Figure A4: Sample Program Description

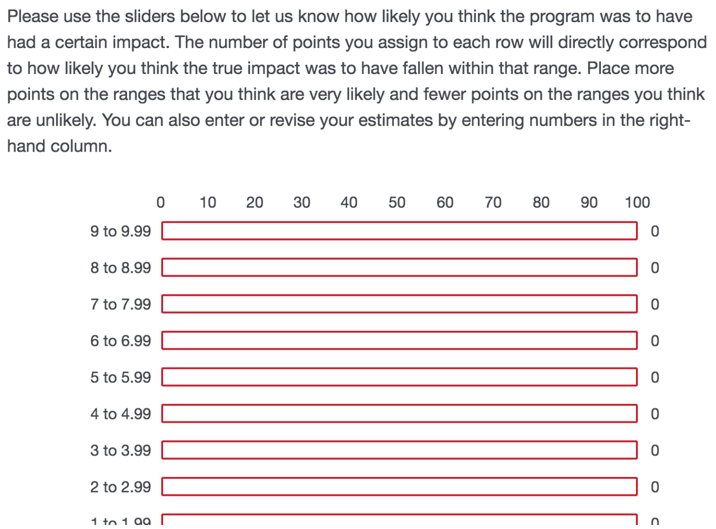


Consider a conditional cash transfer (CCT) program in which a household is provided with the equivalent of $20 USD per month as long as all their children between age 6 and 16 stay in school. The program targets rural areas. Just before the CCT program is implemented, 90% of these children were enrolled in school.

Please provide your best estimate of how much the CCT increased enrolment (in percentage points). Remember that an increase by X percentage points is not the same thing as an increase by X percent!

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 A5: Assigning Likelihoods

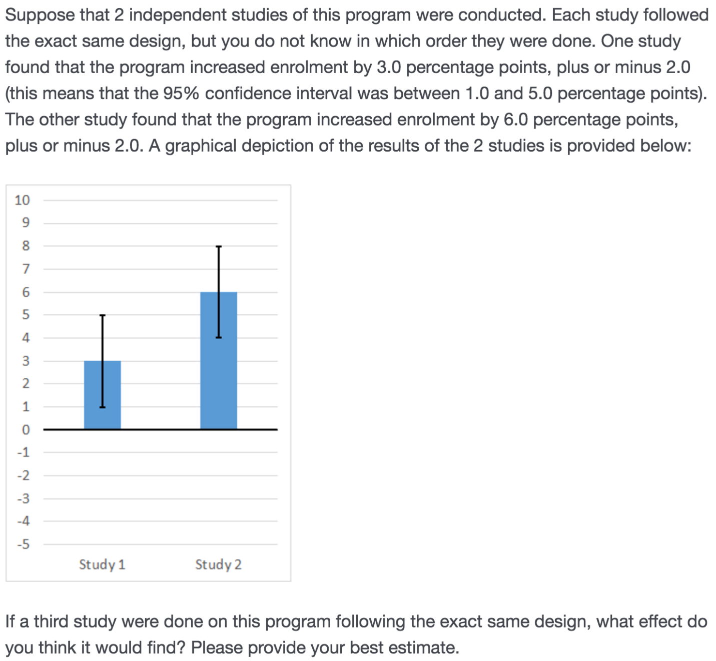


Please use the sliders below to let us know how likely you think the program was to have had a certain impact. The number of points you assign to each row will directly correspond to how likely you think the true impact was to have fallen within that range. Place more points on the ranges that you think are very likely and fewer points on the ranges you think are unlikely. You can also enter or revise your estimates by entering numbers in the right-hand column.

	0 10 20 30 40 50 60 70 80 90 100	
9 to 9.99		0
8 to 8.99		0
7 to 7.99		0
6 to 6.99		0
5 to 5.99		0
4 to 4.99		0
3 to 3.99		0
2 to 2.99		0
1 to 1.99		0

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

Figure A6: Sample New Data

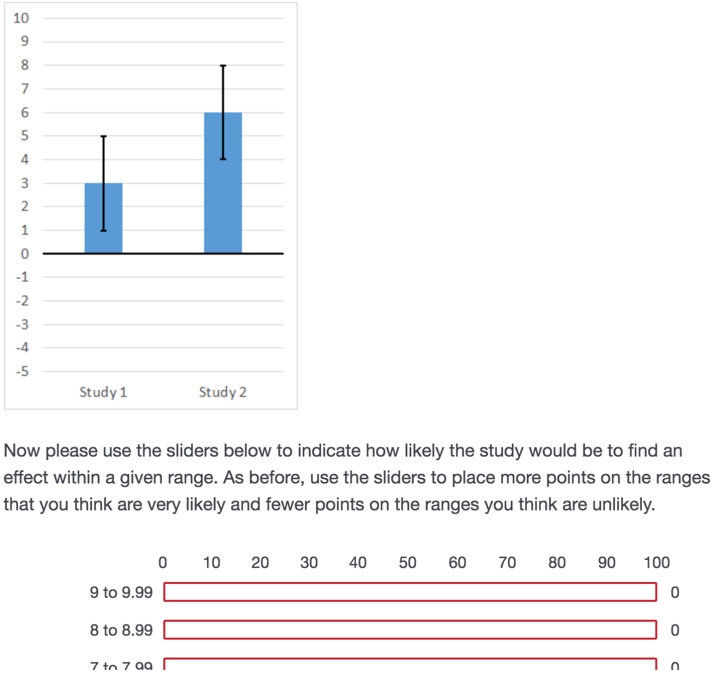


Suppose that 2 independent studies of this program were conducted. Each study followed the exact same design, but you do not know in which order they were done. One study found that the program increased enrolment by 3.0 percentage points, plus or minus 2.0 (this means that the 95% confidence interval 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 A7: Assigning Likelihoods after Viewing New Data

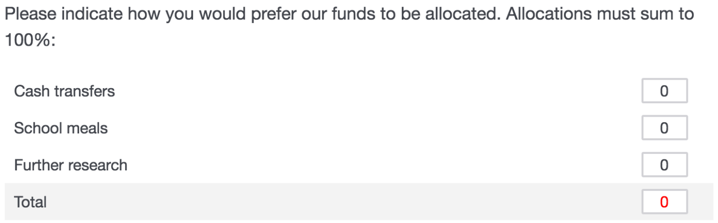


Now please use the sliders below to indicate how likely the study would be to find an effect within a given range. As before, use the sliders to place more points on the ranges that you think are very likely and fewer points on the ranges you think are unlikely.

9 to 9.99	0
8 to 8.99	0
7 to 7.99	0

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

Figure A8: Allocation Question



Please indicate how you would prefer our funds to be allocated. Allocations must sum to 100%:

Cash transfers	0
School meals	0
Further research	0
Total	0

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