# Implementer Desirability Bias in Program Evaluation

Ashish Shenoy[1*] and Travis J. Lybbert[1]

[1]University of California, Davis

[*]Corresponding author. Email: shenoy@ucdavis.edu.

February, 2024

**Abstract**

Development interventions are commonly piloted by organizations with strong community ties. Reminding beneficiaries that a pilot is being evaluated may prompt them to take costly actions that reflect favorably on the implementer. We test for this form of desirability bias in an evaluation of an unsuccessful agricultural extension pilot that ultimately drove treated farmers away from the target crops. Making the evaluation salient during endline data collection led participants to neutralize this negative treatment effect by altering input purchases and cultivation patterns. Participants' desire to support implementers can help explain why some promising pilot results fail to replicate at scale.

**JEL Codes:** L31, C93, O13, O22

# 1 Introduction

Many development programs are initiated as pilots with the intent of scaling if the trial shows success. High-quality implementation at the pilot stage is necessary to ensure a program remains faithful to its intended design. As a result, pilots are commonly run by organizations with strong institutional capacity enabled by a history of local expertise and involvement. However, the intensive community engagement and oversight brought to bear at the pilot stage often cannot be replicated as a program expands in scope. There is growing concern that demonstrations of large program impacts at small scale may have limited external validity as expansion brings in new levels of management and administration.

In this paper, we illustrate how the same features that constrain implementation quality at scale may also bias pilot evaluation, threatening internal validity by overstating true program impacts. Our study takes place in the context of an unsuccessful agricultural intervention to promote smallholder cultivation of pulses in Bihar, India. The policy was piloted in a two-year randomized evaluation by four non-governmental organizations (NGOs) selected for their extensive history of local rural development work in the study area. We uncover evidence that farmers involved in program evaluation take costly actions that make the evaluation appear more favorable to the implementing organizations.

The primary data for this study come from an incentive-compatible elicitation of demand for seeds of the target pulse crops at endline. Actual seed purchases were based on elicited responses, ensuring the decision had real stakes. This exercise was intended to measure farmers' sustained intention to produce pulses after program activities concluded, an explicit goal of the program at the outset. Implementers defined success as an increase in treated farmers' preference for pulse cultivation, resulting in greater seed demand.

To evaluate how implementers' desires affect participant behavior, we experimentally varied the salience of program evaluation during demand elicitation. Specifically, enumerators introduced the elicitation either as an explicit evaluation of the implementer's efforts or more generally as a study of regional attitudes toward pulse cultivation. After this manipulation, elicitation proceeded identically for all participants. Importantly, we ensure the introductory language does not communicate information about product quality by offering a consistent product explicitly sourced and delivered by the local implementer. We interpret differences in participants' willingness-to-pay for pulse seeds by evaluation salience as a reflection of participants' implicit

preferences over the outcome of the evaluation itself.

Increasing the salience of evaluation skews the estimated treatment effect in favor of the implementers. Overall, the two-year intervention actually discouraged pulse cultivation among treated farmers by confirming their belief that growing pulses was not worth the opportunity cost of displacing more lucrative alternatives (see Lybbert et al., 2023, for further details). This belief manifested as 25% lower demand for pulse seeds on average in the incentive-compatible elicitation. However, the negative treatment effect was only observed in elicitations with low evaluation salience, where treated farmers purchased less than half the quantity of their untreated counterparts. By contrast, there was no distinguishable difference in elicitations with high evaluation salience. Making the evaluation salient during data collection obscured evidence of a negative treatment effect.

This shift in seed demand represents costly action taken by study participants. Treated farmers spent an average of Rs. 70 ($1.00) more on pulse seeds in high-salience elicitations. More importantly, seed purchases reflected real cultivation choices over the following crop season. We find a strong, positive correlation between seeds purchased and area planted, with no systematic deviation by salience status. On average, farmers who were reminded of their participation in a program evaluation subsequently altered cultivation on 2% of their cropland in the following season. This reallocation of real on-farm resources, while modest, indicates this bias extends beyond simple survey misreporting or other forms of cheap talk.

Responsiveness to implementer desire can be thought of as a form of Hawthorne effect. Past work has established that subjects in an experiment may alter their behavior when they know they are being monitored or evaluated (see, e.g., Levitt and List, 2011; Friedman and Gokul, 2014; McCambridge et al., 2014). Most relevant to our study, de Quidt et al. (2018) investigate experimenter desirability, whereby participants act in response to researcher objectives, as a possible source of bias. The authors intentionally manipulate beliefs about the experimenter's desires and find the resulting distortions to be modest. We extend this type of work to introduce the possibility that the relevant pressure in program evaluation comes not from the experimenter conducting the evaluation, but rather the implementer being evaluated.

Implementer desirability may influence any participant survey response, but it is especially concerning for complex elicitation methods introduced exclusively to generate evaluation data. In particular, Becker-

DeGroot-Marschack (BDM) elicitation is common in field experiments—including our own study—because it reveals respondents' full demand curve in an incentive-compatible manner (Lusk and Shogren, 2007). This mechanism has been criticized due to concerns about misunderstanding of the dominant strategy (Cason and Plott, 2014; Berry et al., 2020), weak incentives for accuracy (Harrison, 1994), decision fatigue (Brown et al., 2023), and price anchoring (Berry et al., 2020; Mamadehussene and Sguera, 2023). Nevertheless, it empirically matches simpler willingness-to-pay measures that generate less rich data (Berry et al., 2020; Burchardi et al., 2021; Brown et al., 2023). We show that even if an elicitation accurately reveals demand, the demand itself may be influenced by participants' preferences over the outcome of evaluation. This bias is more likely to arise in unnatural exercises that call attention to the research in progress—such as BDM elicitation—than in participants' regular endeavors—such as market purchases.

Our findings more generally contribute to the large literature on reproducibility and policy scaling (see Banerjee et al., 2017; Al-Ubaydli et al., 2019). Across many sectors, programs implemented by NGOs systematically generate greater impact than those run by governments (Vivalt, 2020). In a study closely related to ours, Usmani et al. (2022) highlight the particular importance of prior community engagement in NGO effectiveness. Specifically, Fischer et al. (2019) demonstrate the importance of trust when introducing new market products. Our demand elicitations hold constant trust in the product itself to isolate the effect of participants' preferences over the evaluation outcome.

Community engagement is difficult to replicate at scale (e.g. Dhaliwal and Hanna, 2017; Mitchell et al., 2023), and may therefore threaten the external validity of evaluation results generated from an NGO-implemented pilot. Our research establishes how these factors can also undermine the internal validity of evaluation independent of their role in implementation effectiveness.

Implementer desirability bias can create issues of endogenous participant effort similar to those proposed by Chassang et al. (2012). Many development programs rely on complementary investments from program beneficiaries, and beliefs about implementation quality can alter participants' incentives to invest. For instance, in two field evaluations, Bulte et al. (2014) and Bulte et al. (2023) show how the returns to improved seeds in Tanzania depend crucially on farm labor choices, which are in turn a function of participants' perception of the quality of the seeds being evaluated. We find that participant behavior responds not only

to beliefs about implementation quality, but also preferences over how implementation is evaluated by the researcher.

Heterogeneity in responsiveness to implementer desire would exacerbate external validity concerns regarding favorable selection of sites (Allcott, 2015) or indicators (Saccardo et al., 2023) if NGOs strategically seek out communities where evaluation is likely to show positive impact. This could contribute to the relationship between prior NGO activity and current program effectiveness reported by Usmani et al. (2022), and we report suggestive evidence of such selectivity in our setting.

## 2 Background

The state of Bihar is among the poorest in India with over a third of households below the national poverty line. The population is also predominantly agrarian with just 12 percent residing in urban centers. As a result, Bihar has been a region of focus for rural development programs by both the Government of India and the NGO sector, frequently funded by external donors. There are currently 4,255 registered NGOs and volunteer organizations in the state,[1] the majority of which operate at small scale and rely on heavy engagement with beneficiary communities. We investigate how to translate experience from this type of localized development work into guidance for policy design.

Our study is tied to an agricultural development venture initiated by the Government of India and managed by an international NGO. In 2016, the managing organization enlisted four local Bihari NGOs, each working in a geographically distinct area, to implement a pilot intervention aimed at increasing the production of pulses by farm households. Many households in this region grow small amounts of pulses—primarily pigeon pea—on crop borders or other marginal land for home consumption. The partnership designed and administered an intensive two-year package of input subsidies, agricultural extension services, and marketing support to modernize cropping practices and boost output. This package was piloted in five districts to test whether intensive short-term investment could shift the long-term crop portfolio of participating households.

Implementing partners for this project were selected because of their track record with local development.

[1] Source: NITI Aayog (`https://ngodarpan.gov.in/index.php/home/statewise`) accessed 10/4/2021

All four implementing NGOs had operated in their respective areas for at least ten years prior and had been involved in past initiatives ranging from agriculture to health and nutrition to savings and credit. As a result, local implementers had preexisting relationships with study participants before the inception of the pulses program studied here.

From summer 2017 through spring 2019, implementers procured and distributed certified seeds of improved pulse varieties at subsidized prices, conducted local training and extension sessions to demonstrate best practices, gave individualized feedback to program participants through the cropping season, and assisted with the sale of output. The NGO seed distribution network, which brought in higher quality inputs than were previously available in local markets, remained in place and was opened to all farmers after program activities concluded.

Program activities were carried out in a randomized controlled trial with the intent of using lessons from the pilot to design a statewide policy that could be adopted and run by government agencies. Full program evaluation results are reported by Lybbert et al. (2023). In this paper, we quantify how goodwill toward the implementer may interact with impact evaluation in pilot experimentation.

# 3 Data and Methodology

## 3.1 Data

A key evaluation outcome for the trial was sustained production of pulses after main program activities had concluded. As part of this evaluation, demand for certified pulse seeds was elicited among a sample of study farmers ahead of each planting season in the third year. Specifically, we elicited demand for pigeon pea (*arhar*) and black gram (*urad*) in the summer (*Kharif*) season and for red lentil (*masoor*) in the winter (*Rabi*) season of 2019. All three crops were initially promoted by the pulses program, but second-year implementation predominantly focused on black gram and red lentil. Farmers' willingness-to-pay for these inputs constitute the primary data analyzed in this study.

Each demand elicitation included farmers from the same village who placed orders for certified seed from the NGO supplier through a price list BDM revelation mechanism. For each variety, participants

were given list of possible prices and asked to report the seed quantity they would want to order at each price. At the end of the session, one price at random was selected from the list. Participants agreed to purchase their stated demand at that price, and did not have the option to adjust their quantity demanded after the transaction price was revealed. The purchase transaction ensures that each demand decision was incentive-compatible. Reported demand, in effect, served as the farmer's seed order from the NGO for that season.[2]

After the elicitation exercise, participants were asked a short set of questions related to demographics and intended pulse cultivation. Participants in the winter season were also asked about their subjective perception of whether the pulse program had been beneficial as well as their participation in the pulses pilot program and their past involvement with the local implementer.

### 3.2 Study Design

This study leverages two sources of experimental variation. The primary variation comes from the script introducing the demand elicitation, where we experimentally vary the salience of the program evaluation underway. In half the sessions, designated high-salience, we explicitly announced our relationship with the local NGO and our objective to "evaluate how effective their efforts to promote pulses" had been. In the other half, designated low-salience, we motivated our involvement more generally as a "research project. . . to understand more about pulse production in this region."[3]

Other than this manipulation, demand elicitation proceeded identically for all participants. Notably, the elicitations served as a mechanism for the NGO seed supplier to collect orders, so the partner NGO was identified as the source of seeds and mentioned by name several times throughout elicitation in both high- and low-salience villages. This structure allows us to attribute treatment effects to the salience of evaluation itself, and to rule out that the high-salience script affected participants' beliefs about the products on offer or communicated additional information about the mere involvement of a trusted partner. Salience treatment assignment remained constant within village over the two planting seasons to avoid contamination across

[2]Implementers reported a few isolated instances of farmers refusing purchase at the time of seed delivery, but the vast majority of demand elicitations corresponded to real seed purchases.

[3]The full introductory text is provided in Appendix A. This manipulation and evaluation were pre-registered as a separate trial with the AEA RCT Registry: AEARCTR-0004405.

rounds.

[Table 1 about here.]

Table 1 presents descriptive statistics on participants across salience arms. The first four rows describe the participants themselves, and the next three provide household demographic information. Only the fraction male differs significantly at the 10% level, and a joint F-test fails to reject balanced at the 10% level.

Evaluation salience cross-cuts the second source of variation inherited from treatment assignment in the pilot evaluation. This study includes 94 experimentally treated villages that had received two years of intervention support prior to seed demand elicitation and 53 experimental control villages that did not. Study villages were all identified as potential implementation sites by partner NGOs. Households from these villages were recruited to participate in the program evaluation before initial village treatment assignment, over two years prior to the seed demand elicitations. Both treated and control farmers knew they were participating in an evaluation that may scale up if successful, and as a corollary, control farmers knew they had not been selected to receive the intervention package.

Demand elicitations had a fairly high attrition rate with only two thirds of invited households opting to participate. Participation is uncorrelated with evaluation salience, and the manipulation only took place after recruitment, so attrition does not introduce bias into comparisons across salience arms. However, the choice to participate may have been affected by the prior two years of program participation and therefore may introduce selection bias into comparisons across experimental treatment arms. As a result, our findings in this paper can be interpreted as the causal effect of evaluation salience on the demand of those who participate in the elicitation, but selection into the dataset may differ between treated and untreated groups in the pilot evaluation.[4]

We also elicit demand in 66 villages non-experimentally selected for treatment by the implementing partners alongside pilot evaluation. Appendix A shows there is little difference between elicitation participants in experimentally and non-experimentally treated villages. This non-experimental group allows us to investigate selection into the evaluation sample based on either potential treatment effect or on responsiveness

[4]Roughly half of study farmers chose to attend the seed demand elicitation, balanced across pilot treatment and control. Those who declined are nearly identical to study participants on baseline household characteristics, though they evidently differ in their desire to purchase pulse seeds from the NGO supplier. Further details regarding recruitment and attrition are discussed in Appendix A.

to implementer desirability.

We evaluate the impact of evaluation salience on demand for pulse seeds using the regression

$$Q_{icp} = \beta_1 Treat_i + \beta_2 Salient_i + \beta_3 Salient_i \times Treat_i + \alpha_c + \gamma_p + \delta_{b(i)} + X_i'\sigma + \epsilon_{icp} \quad (1)$$

where $Q_icp$ denotes the quantity demanded by individual $i$ for seeds of crop $c$ at price $p$. $\alpha_c$ and $\gamma_p$ control for crop-specific and price-level demand shifters, respectively. $\delta_{b(i)}$ controls for block-level (sub-district) demand shifters as well as any NGO-specific effects because NGOs are unique by block, and $X_i'\sigma$ controls for participant demographic characteristics.

Coefficients $\beta_1$ and $\beta_2$ in (1) describe how demand differs on average in villages treated in the pilot experiment and those exposed to the high-salience script, respectively. The main coefficient of interest, $\beta_3$, indicates how the salience effect differs between treated and untreated villages. A finding of $\beta_3 \neq 0$ would signify that the estimated treatment effect in the pilot experiment would differ based on whether the fact of evaluation were made salient or not during data collection. Specifically, $\beta_3 > 0$ would correspond to salient evaluations presenting a more favorable view of the intervention in accordance with implementer desirability bias.

# 4 Results

## 4.1 Evaluation Salience and Seed Demand

Our main findings can be observed in the raw (inverse) seed demand curves, plotted for each crop by pilot treatment status and evaluation salience in Figure 1. The dotted lines plot quantity demanded by price among control farmers, the solid lines plot demand among treated farmers, and the gap between them represents the measured treatment effect. Average demand in low-salience elicitations are plotted in black and high-salience in red.

Treated farmers' seed demand is consistently below that of untreated farmers in low-salience elicitations, corresponding to their negative impression of pulse cultivation through two years of intervention. However,

this gap narrows, and in the case of red lentil actually reverses, in elicitations where the impact evaluation was made salient. The salience effect is most apparent with black gram and red lentils, the two crops that were the primary focus of second-year program activities. By contrast, for pigeon peas, which farmers were most likely to have already been growing pre-intervention, salience lowers demand overall but the difference in effect by treatment status is modest.

[Figure 1 about here.]

Table 2 confirms these patterns in regression estimates of differences in seed quantity demanded by treatment assignment in the pilot experiment and by evaluation salience in the demand elicitation. All standard errors are clustered at the village level. We report p-values from randomization inference over 1,000 iterations reassigning village-level salience status in square brackets.

[Table 2 about here.]

The first row of Column 1 aligns with the main program evaluation conclusion that promotion activities actually lowered input demand among treated farmers. This result arises because two years of experience confirmed treated farmers' prior belief that pulses are no more profitable than the crops they would displace, even with improved agronomic practices, higher quality seeds, and technical support. Relative to the mean, treatment lowered pulse demand by nearly 25% on average.[5]

The primary finding of this paper is that the effect of evaluation salience varies with treatment assignment in the pilot experiment. As shown in the third row of Column 1, making the impact evaluation salient lowers demand on average. Column 2 breaks this effect down by treatment assignment according to the regression specification in (1). The second row reveals a stark difference in response between those in treated and untreated villages.

The positive sign of the coefficient on the interaction term ($\beta_3$) indicates treated farmers signal greater demand when evaluation is salient relative to untreated farmers. This behavior would support implementing partners' desire to demonstrate the success of their intervention. The magnitude of the coefficient—95% as

[5]Lybbert et al. (2023) show the demand elicitation impacts to be consistent with other post-intervention indicators that treated farmers ceased pulse cultivation once subsidies expired, and not attributable to differential attendance at demand elicitations or to buildup of stocks during the intervention period. However, a causal interpretation of the treatment effect is not necessary for the discussion regarding implementer desirability in this study.

large as the negative effect in the first row—suggests evaluation salience distorts results by enough to nearly erase the decline in demand caused by treatment. A joint test fails to reject that the estimated treatment effect is statistically distinguishable from zero among high-salience participants at the 10% level. Crop-specific regression results, presented in Appendix Figure S2, reveal this pattern to be qualitatively present for all three crops, but strongest for black gram and red lentil.

These results are consistent with farmers adjusting their behavior to satisfy implementers' desire to demonstrate effectiveness when reminded of the program evaluation. Mentioning the evaluation lowers seed demand on average as participants are reminded of the unsuccessful intervention over the prior two years. However, treated farmers appear to resist this pressure more than untreated farmers, making the intervention appear to have had a more positive (i.e. less negative) impact on their preference to grow pulses. Conversely, those in the experimental control group may seek to demonstrate an unmet need for further NGO involvement by suppressing their desire to test out high-quality seeds without NGO support. In either case, the altered valuation by study participants leads evaluation to look more favorable to the implementer.

For comparison, Column 3 of Table 2 reports farmers' stated belief about whether the pulses treatment package was beneficial to those who received it. Responses to this question, asked only in the Winter session after seed demand elicitation had concluded, are highly favorable, with three quarters of participants claiming to believe the program was beneficial. There is no statistically significant variation in this rate with treatment assignment, and no evidence of unfavorable attitudes among treated farmers. However, responses to this question are cheap talk and therefore less informative than real-stakes purchase decisions.

## 4.2 Magnitude of Implementer Desirability Effect

Implementer desirability bias manifests as a real shift in agricultural portfolios. The estimated difference of a half kilogram of purchased pulse seed corresponds to spending about Rs. 70, or $1.00, more at supplier prices. While this expenditure by itself may be small, program implementers were active in ensuring farmers planted and cultivated what they ordered. As a result, the main cost to participants came in the acreage and effort they subsequently devoted to pulse farming.

Purchase quantity and area cultivated were documented by the NGO supplier for all farmers from villages that had previously received the pulses intervention. Table 3 reports the relationship between these outcomes in administrative records. The table shows a strong, positive correlation with a regression R-squared of between 0.6 and 0.8. Importantly, there is no systematic pattern in deviations from this relationship by evaluation salience. After controlling for seed quantity, the effect of evaluation salience on acreage is quantitatively small and statistically indistinguishable from zero. This evidence indicates that elicited seed demand corresponds to real differences in area planted, and not merely performative purchases made at low cost and subsequently discarded or given to another farmer.

[Table 3 about here.]

For black gram and red lentil—the two crops where evaluation salience has the greatest impact—one kilogram of seeds purchased corresponds to roughly 0.07 acres cultivated. Multiplying this by the effect on seed purchases, we estimate implementer desirability bias led farmers to reallocate around 0.03 crop acres, or 2% of their total landholdings. While this represents a small portion of their total agricultural portfolio, the investment of land—and associated labor and other inputs—nevertheless reflects a real and sustained contribution of costly on-farm resources when program evaluation is made salient.

This quantification must be interpreted with two caveats. First, administrative records were only kept in villages previously treated in the pilot evaluation, and no comparable records exist for untreated villages. This omission, while unfortunate, is not so concerning because untreated farmers, if anything, reduced their seed order in response to implementer desirability. The risk of performative overpurchasing during demand elicitation—i.e. buying excess unplanted seeds—lies with treated farmers, and it is exactly this group for whom we have data.

Second, NGO transaction records cannot be linked to individuals in the demand elicitation, so salience treatment status is assigned at the village level. Experimental elicitations accounted for only 35–40% of purchases represented in Table 3, with the remainder being regular orders from non-participants who were not exposed to the salience script. However, even multiplying the estimated salience effect by three would not offset the implied acreage expansion, though this calculation is noisy. Evidence suggests purchasing behavior translated into a modest but real reallocation of resources in accordance with implementer

desirability.

### 4.3 Favorable Site Selection

Responsiveness to implementer desirability may offer a dimension on which organizations seeking favorable evaluations select pilot locations. In our study, evaluation was initiated with the explicit intent of scaling up under the leadership of the implementing partners if the intervention package generated a sustained increase in pulse production. We investigate the possibility that study sites were strategically selected in response to evaluation incentives by comparing seed demand in the experimental sample against the 66 villages non-experimentally selected for treatment by the implementing NGOs.

Farmers treated outside the impact evaluation appear less responsive to implementer desirability and less favorable toward the program overall. Columns 4 and 5 of Table 2 reproduce Columns 2 and 3 with those non-experimentally selected by implementing partners. Seed demand in the non-experimentally treated group is lower than in the experimentally treated group. Notably, the demand response in response to the high-salience script, while positive, only offsets half the negative treatment effect.

While neither of these differences are statistically distinguishable from the experimentally treated sample on their own, the net treatment effect following the high-salience script remains negative and statistically distinguishable from zero when comparing non-experimentally treated farmers to the experimental control. Stated beliefs about program benefits are also lower in this group than in the experimental sample. Together, the evidence suggests farmers treated in the experimental evaluation are more likely to demonstrate positive treatment effects, especially in response to implementer desirability, consistent with selection on evaluation favorability.[6]

Even this level of selectivity may be attenuated due to the timing of rollout. The pulses program was initially envisioned as a proof-of-concept, and the randomized evaluation component was subsequently introduced after treatment had already been promised in around ten villages.[7] Therefore, non-experimentally treated farmers include both those selected for piloting pre-randomization—which we would expect to be

[6] In Appendix B we explore participant-level heterogeneity in involvement with the implementing NGO. This is a self-reported measure that responds endogenously to the salience script, but evidence shows little variation along this margin.

[7] Unfortunately the administrative record-keeping lacks sufficient detail to explore heterogeneity in the timing of non-experimental village enrollment.

the most favorable sites—as well as those enrolled after experimental activities were underway—which we would expect to be less favorable.

It should be noted participant selection for demand elicitations differed slightly in non-experimental villages. In experimental villages, farmers were selected at random out of the group that had expressed interest in pulses prior to the first season of program implementation two years prior, some of whom may have subsequently opted out of treatment. No such roster existed in non-experimental villages, so farmers were invited at random out of those who participated in the first implementation season. Table S4 reveals minor differences in participant characteristics between experimentally and non-experimentally treated villages, but no systematic selection patterns.

# 5 Discussion

When evaluating a program run by a sympathetic implementer, the salience of the evaluation itself can positively bias the estimated treatment effect. We establish this fact in the context of an unsuccessful pilot agricultural intervention that actually reduced farmer demand for the target crops but was implemented by organizations with strong community ties. When the evaluation was made salient, participants altered their behavior in a real-stakes demand elicitation with seemingly binding input decisions for the coming crop season. Making the evaluation salient effectively closed the gap between treatment and control, masking evidence of the negative program impact on seed demand.

This outcome is consistent with participants adjusting seed purchases to align with the desires of the program implementer. Our study involves four implementers, each of which have a long history and are viewed favorably in their respective regions of operation. Nevertheless, there is heterogeneity across implementers in responsiveness to implementer desirability, plotted in Figure S3. With only four data points, we hesitate to speculate on the specific aspects of implementer identity that influence participants' preferences over evaluation outcomes. Applying our methodology to program evaluation involving a larger number of implementing organizations can help shed light on the exact characteristics most likely to trigger desirability bias.

Motivation to support favorable evaluation may be either forward-looking or backward-looking. One

possibility is that program beneficiaries understand that NGOs' funding and continued operation depend crucially on their ability to demonstrate success, and strategically act to deliver positive evaluation results. In this case, the costly actions taken by evaluation participants can be seen as investments in anticipation of a future stream of benefits that will remain intact as long as the implementing NGO remains active in the area.

Alternatively, invoking program evaluation at the start of data collection may induce feelings of reciprocity toward the implementer. Participants may wish to reward the implementer in exchange for the benefits of the program itself, goodwill generated by the effort put in by implementing agents, or other aspects of service delivery. Whatever the catalyst, beneficiaries of development programs can reciprocate at the evaluation stage by behaving in accordance with program goals. Our investigation focuses on identifying and quantifying the size of implementer desirability bias, but isolating the exact channel through which it operates remains an open topic for future research.

The bias identified in this study represents a specific threat to reproducibility with clear implications for translating experience from pilot programs into broader policy lessons. When introducing new development initiatives, it is common practice to partner with established organizations that have strong community ties to leverage their local knowledge and institutional capacity. However, our findings suggest the features that enable successful implementation are precisely those that can undermine accurate evaluation. When faced with the prospect of evaluation, beneficiaries sympathetic to the implementer may take costly actions that help the implementer achieve its goal of demonstrating success. As a consequence, pilot evaluation can uncover misleadingly optimistic results that overstate a policy's true impacts.

# References


**Al-Ubaydli, Omar, John A List, and Dana Suskind**, "The Science of Using Science: Towards an Understanding of the Threats to Scaling Experiments," Working Paper 25848, National Bureau of Economic Research May 2019.

**Allcott, Hunt**, "Site Selection Bias in Program Evaluation," *The Quarterly Journal of Economics*, 2015, *130* (3), 1117–1165.

**Banerjee, A.V., S. Chassang, and E. Snowberg**, "Decision Theoretic Approaches to Experiment Design and External Validity," in Abhijit Vinayak Banerjee and Esther Duflo, eds., *Handbook of Field Experiments*, Vol. 1 of *Handbook of Economic Field Experiments*, North-Holland, 2017, pp. 141–174.

**Berry, James, Greg Fischer, and Raymond Guiteras**, "Eliciting and Utilizing Willingness to Pay: Evidence from Field Trials in Northern Ghana," *Journal of Political Economy*, 2020, *128* (4), 1436–73.

**Brown, Alexander, Jinliang Liu, and Michael Tsoi**, "Is There a Better Way to Elicit Valuations than the BDM?," Working Paper 4476764, Social Science Research Network 2023.

**Bulte, Erwin, Gonne Beekman, Salvatore Di Falco, Joseph Hella, and Pan Lei**, "Behavioral Responses and the Impact of New Agricultural Technologies: Evidence from a Double-blind Field Experiment in Tanzania," *American Journal of Agricultural Economics*, 2014, *96* (3), 813–830.

**_ , Salvatore Di Falco, Menale Kassie, and Xavier Vollenweider**, "Low-Quality Seeds, Labor Supply and Economic Returns: Experimental Evidence from Tanzania," *The Review of Economics and Statistics*, 2023, *forthcoming*.

**Burchardi, Konrad B., Jonathan de Quidt, Selim Gulesci, Benedetta Lerva, and Stefano Tripodi**, "Testing willingness to pay elicitation mechanisms in the field: Evidence from Uganda," *Journal of Development Economics*, 2021, *152* (Sep), 102701.

**Cason, Timothy N. and Charles R. Plott**, "Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing," *Journal of Political Economy*, 2014, *122* (6), 1235–70.

**Chassang, Sylvain, Gerard Padró I Miquel, and Erik Snowberg**, "Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments," *American Economic Review*, 2012, *102* (4), 1279–1309.

**de Quidt, Jonathan, Johannes Haushofer, and Christopher Roth**, "Measuring and Bounding Experimenter Demand," *American Economic Review*, November 2018, *108* (11), 3266–3302.

**Dhaliwal, Iqbal and Rema Hanna**, "The devil is in the details: The successes and limitations of bureaucratic reform in India," *Journal of Development Economics*, 2017, *124*, 1–21.

**Fischer, Greg, Dean Karlan, Margaret McConnell, and Pia Raffler**, "Short-term subsidies and seller type: A health products experiment in Uganda," *Journal of Development Economics*, 2019, *137*, 110–124.

**Friedman, Jed and Brinda Gokul**, "Quantifying the Hawthorne Effect," Technical Report October 16, World Bank Blogs 2014.

**Harrison, Glenn**, "Expected Utility Theory and the Experimentalists," *Empirical Economics*, 1994, *19* (2), 223–53.

**Levitt, Steven D. and John A. List**, "Was There Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments," *American Economic Journal: Applied Economics*, 2011, *3* (1), 224–38.

**Lusk, Jayson L. and Jason F. Shogren**, *Experimental Auctions: Methods and Applications in Economic and Marketing Research*, Cambridge, UK: Cambridge University Press, 2007.

**Lybbert, Travis, Ashish Shenoy, Tomoe Bourdier, and Caitlin Kieran**, "Striving to Revive Pulses in India with Extension, Input Subsidies, and Output Price Supports," *American Journal of Agricultural Economics*, 2023, *forthcoming*.

**Mamadehussene, Samir and Francesco Sguera**, "On the Reliability of the BDM Mechanism," *Management Science*, 2023, *69* (2), 1166—-79.

**McCambridge, Jim, John Witton, and Diana R. Elbourne**, "Hawthorne effect: new concepts are needed to study research participation effects," *Journal of Clinical Epidemiology*, 2014, *67* (3), 267–77.

**Mitchell, Harrison, A. Mushfiq Mobarak, Karim Naguib, Maira Rei ao, and Ashish Shenoy**, "Delegation Risk and Implementation at Scale: Evidence from a Migration Loan Program in Bangladesh," Unpublished manuscript 2023.

**Saccardo, Silvia, Hengchen Dai, Maria Han, Naveen Raja, Sitaram Vangala, and Daniel Croymans**, "Assessing Nudge Scalability," Working Paper 3971192, Social Science Research Network 2023.

**Usmani, Faraz, Marc Jeuland, and Subhrendu K. Pattanayak**, "NGOs and the Effectiveness of Interventions," *Review of Economics and Statistics*, 2022, *forthcoming.*

**Vivalt, Eva**, "How Much Can We Generalize From Impact Evaluations?," *Journal of the European Economic Association*, 2020, *18* (6), 3045–3089.

Table 1: Participant Characteristics by Treatment Assignment

| | Evaluation Salience | | |
|---|---|---|---|
| | Low | High | Difference |
| Male | 0.87 | 0.80 | −0.07 |
| | (0.34) | (0.40) | [0.10] |
| Age | 47.23 | 47.87 | 0.64 |
| | (16.65) | (17.10) | [0.66] |
| Primary School | 0.64 | 0.65 | 0.01 |
| | (0.48) | (0.48) | [0.90] |
| Secondary School | 0.46 | 0.50 | 0.04 |
| | (0.50) | (0.50) | [0.55] |
| HH Size | 7.26 | 7.22 | −0.04 |
| | (3.69) | (3.84) | [0.90] |
| SC/ST | 0.17 | 0.13 | −0.04 |
| | (0.37) | (0.33) | [0.32] |
| Acres Owned | 1.74 | 1.66 | −0.08 |
| | (1.96) | (1.80) | [0.67] |
| Joint Significance | | | 0.18 |
| Participation Rate | 0.67 | 0.63 | −0.04 |
| Participants | 333 | 372 | |
| Villages | 69 | 78 | |

Notes: Group averages with standard deviations in parentheses and p-value of difference in square brackets. Rows correspond to fraction male, participant age, primary school completion, secondary school completion, household size, fraction belonging to a schedule caste or scheduled tribe, and land area owned by household at pilot baseline. Participation rate reflects fraction of those invited who appeared at either demand elicitation.

Table 2: Effect of Evaluation Salience on Seed Quantity Demanded

| | Experimental | | | Non-Experimental | |
|---|---|---|---|---|---|
| | Seed Demand | | Stated Belief | Seed Demand | Stated Belief |
| | (1) | (2) | (3) | (4) | (5) |
| Treated | -0.260 | -0.531 | 0.132 | -0.728 | -0.081 |
| | (0.12) | (0.19) | (0.08) | (0.20) | (0.11) |
| Salient × Treated | | 0.506 | -0.152 | 0.385 | 0.092 |
| | | (0.23) | (0.11) | (0.22) | (0.14) |
| | | [0.04] | [0.22] | [0.12] | [0.58] |
| Salient | -0.171 | -0.491 | 0.059 | -0.551 | 0.075 |
| | (0.10) | (0.19) | (0.08) | (0.18) | (0.09) |
| | [0.13] | [0.01] | [0.55] | [0.00] | [0.49] |
| Variable Mean | 1.12 | 1.12 | 0.74 | 1.15 | 0.70 |
| Salient Treat Effect (p-val.) | | 0.85 | 0.79 | 0.01 | 0.91 |
| R-Squared | 0.16 | 0.16 | 0.08 | 0.18 | 0.10 |
| Observations | 7390 | 7390 | 451 | 4700 | 266 |

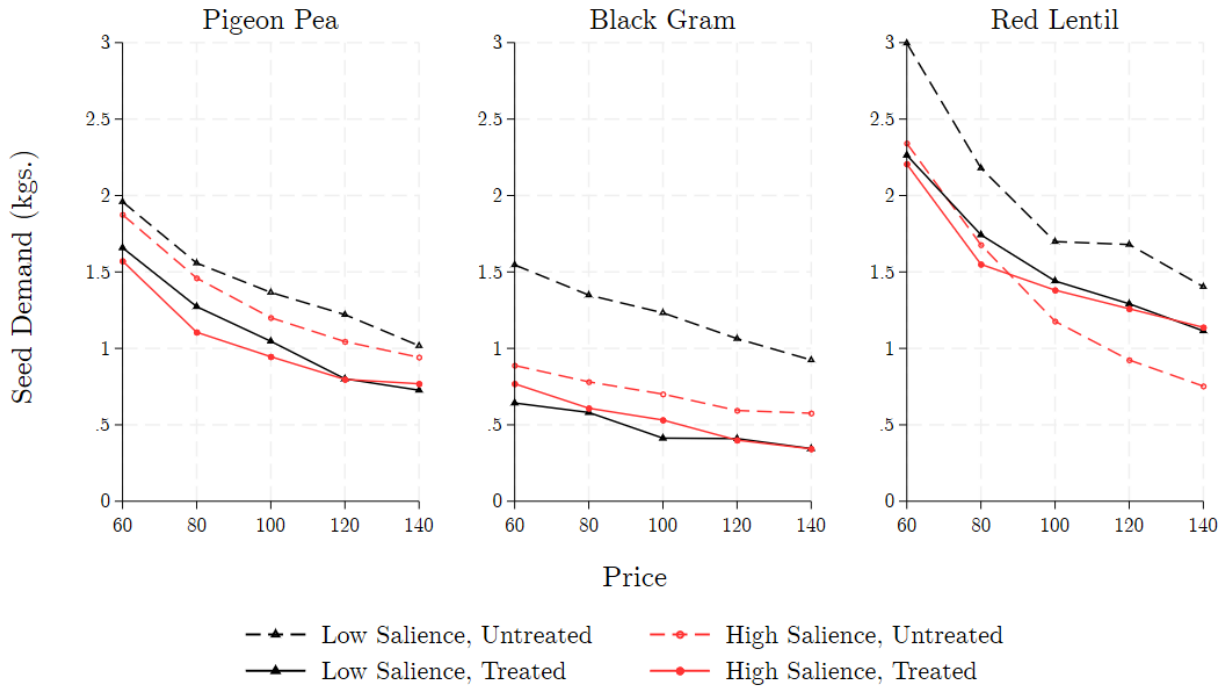
Notes: Outcome is pulse seed quantity demanded. Treated: Village received pulse program treatment in prior two years. Salient: High-salience script in elicitation. Pulse Program: Participant self-identifies as beneficiary of pulse program. Prior NGO: Participant self-identifies as beneficiary of previous NGO programs. Salient Treat Effect: p-value from test of sum of coefficients on Treated and Salient × Treated. All regressions include crop and price fixed effects, block (sub-district) fixed effects, and participant demographic controls. Columns (3) and (4) restrict to those that self-reported program participation during the red lentil (winter) elicitation. Standard errors clustered by village in parentheses; p-values from randomization inference over 1,000 re-draws of village-level salience treatment status in square brackets.

Table 3: Seed Purchases and Cultivated Area (Treated Farmers)

| | All | Pigeon Pea | Black Gram | Red Lentil |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Seed Quantity | 0.078 | 0.163 | 0.073 | 0.065 |
| | (0.00) | (0.01) | (0.00) | (0.00) |
| Salience Treatment | -0.004 | -0.010 | -0.011 | -0.006 |
| | (0.01) | (0.01) | (0.01) | (0.01) |
| Mean Seed (kg.) | 1.61 | 0.97 | 0.97 | 1.92 |
| Mean Area (acre) | 0.15 | 0.17 | 0.09 | 0.15 |
| R-Squared | 0.60 | 0.79 | 0.75 | 0.61 |
| Observations | 11087 | 2402 | 1138 | 7547 |

Notes: Outcome is acres devoted to pulse crop. Input quantity is measured in kgs. of seed purchased from NGO supplier. Column (1) uses data from all crops and includes crop fixed effects; Columns (2)–(4) present results by crop. Data only include farmers from villages previously treated in pilot evaluation because records were not kept in previously untreated villages. Data are from the two seasons of experimental demand elicitation and include additional purchase and planting from those not involved in demand elicitation. Full data are presented in Appendix Figure S1. Standard errors clustered by village in parentheses.

Figure 1: Inverse Demand Curves by Crop and by Experimental Status




Notes: Average seed demand among farmers by pilot treatment status and evaluation salience at each price level for each crop.

## Acknowledgments and Disclosure Statement

We are indebted to the study and survey participants for generously giving their time and, at early stages, sharing their insights in focus group settings. We are grateful to the Aga Khan Foundation, the Aga Khan Rural Support Programme, Kaushalya Foundation, Nav Jagrati, and SSEVS for local support, coordination, and direction. We thank Komal Jain, Nandish Kenia, and members of NEERMAN for design input and data collection; Tomoé Bourdier, Caitlin Kieran, and Stamatina Kotsakou for research assistance; and Tony Cavalieri, Mariana Kim, and Marcella McClatchey for policy coordination, feedback, and financial and logistical support. We appreciate the support and contributions of NITI-Aayog from the conception to the completion of this study, including Ramesh Chand and his team. We thank seminar audiences at the UC Davis, UC San Francisco, UC Santa Cruz, Northwestern University, and Y-RISE for helpful feedback.

Data collection was funded by the Bill and Melinda Gates Foundation. Evaluation funding included two and a half months of summer salary each for authors Lybbert and Shenoy. Authors declare we have no further conflicts of interest. No institution had the right to review results before publication.

All data collection was approved by the University of California, Davis IRB. This study was pre-registered at the AEA RCT registry under AEARCTR-0004405, and the main pilot evaluation under AEARCTR-0003872.

# Supplementary Appendix for

# "Implementer Desirability Bias in Program Evaluation"

# For Online Publication Only

## A Experiment Details

### A.1 Participant Recruitment

Study participants from the pilot experiment—both treated and control—were selected based on interest in pulses prior to randomization, two years before the input demand elicitation. Before the pilot intervention began, implementing partners held kickoff meetings in each experimental village to advertise the program and identify interested farmers. Seven–eight attendees from each kickoff meeting were selected at random for surveying during the intervention period, and they constitute the set of invited participants in seed demand elicitation from experimental villages. In non-experimentally treated villages, we attempt to recreate this selection procedure as closely as possible by sampling at random out of the initial set of farmers engaged by the NGO when they first began program activities in the village.

In practice, only two thirds of invited farmers actually participated in either demand elicitation, with little difference between control, experimental treatment, and non-experimental treatment. Among those invited, 50% participated in the summer (Kharif) session and 42% in the subsequent winter (Rabi) session. Roughly a third of invitees for the summer session and half in the winter session declined due to lack of interest or other local engagements. The rest either could not be reached or were unavailable on the day of the elicitation for reasons such as travel outside the village. Those who were invited and participated constitute the sample used in this study.

[Table S1 about here.]

[Table S2 about here.]

The top panels of Tables S1 and S2 report the baseline characteristics of invited households by participation status in each demand elicitations. Participants and non-participants are nearly identical on baseline household characteristics, though the obviously differ in their desire to purchase pulse seeds. Participation was determined before reading the evaluation salience script, so attrition does not affect the causal interpretation of the effect of evaluation salience.

However, attrition should be taken into account when drawing conclusions about who responds to evaluation salience in their input demand.

To reach the target of 8–10 participants in each demand elicitation session, we requested village leaders to invite additional farmers that were interested in purchasing pulse seeds. The bottom panels of Tables S1 and S2 compare this supplemental sample to the main sample on the limited set of characteristics we collected data on during demand elicitation. Supplemental participants are typically older and more male than invited participants, and the difference is statistically significant at the 1% level.

More worryingly, field reports indicate many of these supplemental participants came from neighboring villages so we cannot accurately determine their treatment assignment or even verify their participation in the pilot experiment. This fact is also reflected in their substantially lower rate of self-identifying as a participant in the pulses program, shown in Table S2. As a result, we exclude them from the main analysis.

[Table S3 about here.]

Table S3 presents regression results from this supplemental sample as the counterpart to Table 2. Estimated effects are in the same direction as in the main sample, but not as strong. This result is consistent with inaccurately recorded participation in the pilot experiment, which would both weaken the effect of evaluation salience and introduce attenuation bias into regressors based on pilot treatment status.

## A.2 Village Selection

This study takes place in villages that participated in a pilot evaluation of an initiative to promote pulse production. The evaluation initially comprised 158 villages, of which 99 were experimentally assigned to receive the intervention package and the other 59 were assigned to control. The input demand elicitation took place after two years of intervention. At the time of elicitation, we included farmers from an additional 70 villages selected non-experimentally for treatment by the implementing partners. Farmers from 94 of the 99 experimentally treated villages, 53 of the 59 control villages, and 66 of the 70 non-experimentally treated villages consented to participate in demand elicitations.

Note that the comparison between the experimental control group and those non-experimentally treated is not random. The ideal experiment would be to randomly un-treat a portion of the non-experimental villages and compare treated to untreated within this sample. This comparison is unavailable, so we use the experimental control group as a counterfactual for those non-experimentally treated. Therefore, the results of this comparison should be considered suggestive and not experimentally identified.

Participant selection for the demand elicitation in non-experimentally treated villages differed from that of experimentally treated villages. Attendance records from initial kickoff meetings in non-experimental villages two years prior were not retained by implementers, so participants were recruited from among the set of farmers that participated in the first season of first season of program implementation.

[Table S4 about here.]

Table S4 reports how differential selection affected auction participants in non-experimental villages. Participants differ on a few observable characteristics, but there is no systematic selection pattern. Notably, participation rates are equal between groups and a joint test of significance fails to reject equality between experimentally and non-experimentally treated participants at the 10% level.

[Table S5 about here.]

Table S5 reports differences in quantity demanded and self-reported measures of NGO engagement between these groups. Non-experimentally treated farmers are less responsive to evaluation salience in the demand elicitation and less likely to state the belief that the pulses program was beneficial, though the latter difference disappears when evaluation is made salient. Both differences make outcomes appear more favorable to the implementing NGOs in the experimental evaluation.

## A.3 Salience Script and Demand Elicitation

Enumerators introduced themselves before the demand elicitation exercise using one of two scripts. In the high-salience version, enumerators read the following paragraph:

> "Your participation in this auction and the survey is part of an evaluation project. We are here because of our partnership with [PARTNER ORGANIZATION] to help evaluate how effective their efforts to promote pulses are. To do so, we want to understand how beneficial you think pulse production would be to you as a farmer after their work in this region."

In the low-salience version, enumerators introduced themselves with the following paragraph:

> "Your participation in this auction and the survey is part of a research project. As such, we are here as a research team, not a sales team. We are not here to promote any kind of pulses. We simply want to understand how beneficial you think pulse production would be to you as a farmer and understand more about pulse production in this region."

If participants asked specifically about the partner NGO or the pulses program during low-salience sessions, enumerators were instructed to provide the following response:

> "We are aware of their activities, but this exercise is designed to learn about attitudes to pulses in this region overall. We are visiting several villages in this area, including many where [PARTNER ORGANIZATION] is not operating."

Other than this difference, demand elicitation proceeded identically in all villages.

In each elicitation, participants were given a list of possible prices for certified varieties of pigeon pea and black gram in the summer (Kharif) session, and of red lentil in the winter (Rabi) session. They reported quantity demanded at each possible price, and then one price was selected at random for actual purchase. To ensure incentive-compatibility, participants could not adjust their quantity demanded after the price was announced.

The maximum price in each elicitation was the prevailing supplier price. Above this level, demand would rationally have been zero because participants always have the option to purchase seeds directly from the supplier outside our elicitation. Hence, the elicitation reflects demand when purchasing seeds at a discount relative to their outside option, and we cannot observe hypothetical demand at higher prices.

Prior to demand elicitation, participants played two practice rounds to build familiarity with the mechanism. In the first practice round, they were given a participation fee and could opt to purchase sweets from the enumerators. In the second practice round, they stated hypothetical seed demand, drew a hypothetical transaction price, and were told the quantity of seed they would have received were it the real elicitation. In field testing, we found this second practice round was necessary to ensure participants understood they would not be able to adjust their quantity after observing the real transaction price.

## A.4 Seed Purchases and Planting

The demand elicitation was designed to measure participants' valuation for certified pulse seeds. The difference between treated and untreated farmers in the pilot experiment reflects changes in valuation induced by the prior two years of learning-by-doing, and variation in the salience language tests whether this valuation was affected by preferences over the outcome of the evaluation itself. This interpretation would be confounded if the salience script also communicated information about seed quality or NGO involvement in general. In particular, seeds sourced by the NGO were more reliable than those available in local markets, so it was important to ensure that the high-salience script did not add certainty about the source of the seeds on offer. To avoid this possibility, all demand elicitations provided coupons for

the specified price and quantity of certified seeds to purchase directly from the NGO supplier. In this way, we hold constant participants' beliefs about the identity of the product being sold and the involvement of the NGO in the sale, and isolate variation in participants' personal seed valuation.

Seed delivery and payment took place several days after the demand elicitation exercise, so coupon details were also shared directly with the supplier in case farmers lost their coupon. At the time of delivery, participants could not alter their quantity demanded at the discounted coupon price. However, they had the option to purchase a different amount at the supplier price (foregoing the experimental discount) or to purchase nothing at all. Suppliers unfortunately did not keep detailed records of coupon redemption, but field reports indicate there were only a few isolated instances where farmers refused purchase and no known instances of farmers purchasing a different amount at the supplier price. Therefore, we interpret experimental responses as farmers' true seed demand at the time of the elicitation.

After delivering seeds, implementing NGOs were active in ensuring farmers planted what they purchased. Seed quantity and area cultivated were documented in administrative records for all farmers from villages that received the pulses intervention. Figure S1 plots the administrative data by crop, corresponding to the regression estimates reported in Table 3. The figure reveals a strong, positive relationship between for each crop in the demand elicitation. Moreover, the acreage of farmers in both high-salience and low-salience villages are evenly distributed around the trend line. Unfortunately, these records cannot be linked to participants in the demand elicitation, and no comparable records exist for farmers from untreated villages in the pilot evaluation.

[Figure S1 about here.]

Table S1: Participation in the Summer (Kharif) Demand Elicitation

| Household Characteristics | Main Sample | Declined | Unavailable | |
|---|---|---|---|---|
| Farmer Male | 0.86 | 0.84 | 0.85 | |
| | (0.35) | (0.37) | (0.36) | |
| Farmer Age | 48.92 | 48.20 | 48.73 | |
| | (16.34) | (15.68) | (15.24) | |
| Farmer Primary Sch. | 0.63 | 0.59 | 0.66 | |
| | (0.48) | (0.49) | (0.47) | |
| Farmer Secondary Sch. | 0.45 | 0.44 | 0.51 | |
| | (0.50) | (0.50) | (0.50) | |
| HH Size | 7.13 | 7.22 | 7.09 | |
| | (3.55) | (3.78) | (4.29) | |
| SC/ST | 0.14 | 0.17 | 0.13 | |
| | (0.35) | (0.38) | (0.34) | |
| Acres Owned | 1.61 | 1.57 | 1.66 | |
| | (1.74) | (1.96) | (1.75) | |
| Joint Difference | | 0.92 | 0.97 | |
| Participants | 534 | 376 | 171 | |
| Participant Characteristics | Main Sample | | | Supplemental |
| Male | 0.83 | | | 0.91 |
| | (0.37) | | | (0.29) |
| Age | 46.50 | | | 47.08 |
| | (16.96) | | | (42.23) |
| Primary School | 0.66 | | | 0.60 |
| | (0.48) | | | (0.49) |
| Secondary School | 0.49 | | | 0.42 |
| | (0.50) | | | (0.49) |
| HH Size | 7.13 | | | 7.84 |
| | (3.55) | | | (3.98) |
| SC/ST | 0.14 | | | 0.13 |
| | (0.35) | | | (0.33) |
| Saved Seeds | 1.97 | | | 1.94 |
| | (5.22) | | | (4.56) |
| Joint Difference | | | | 0.00 |
| Participants | 534 | | | 702 |

Notes: Group averages with standard deviations in parentheses. Top panel describes household characteristics by participation among invited farmers. Rows correspond to fraction male, farmer age, primary school completion, secondary school completion, household size, fraction belonging to a schedule caste or scheduled tribe, and land area owned by household at pilot baseline. Bottom panel describes participant characteristics among main and supplemental sample. Rows correspond to fraction male, participant age, primary school completion, secondary school completion, household size, fraction belonging to a schedule caste or scheduled tribe, and self-reported stock of saved pulse seeds for planting.

Table S2: Participation in the Summer Winter (Rabi) Demand Elicitation

| Household Characteristics | Main Sample | Declined | Unavailable | |
|---|---|---|---|---|
| Farmer Male | 0.85 | 0.84 | 0.88 | |
| | (0.35) | (0.37) | (0.33) | |
| Farmer Age | 49.93 | 47.91 | 45.00 | |
| | (16.07) | (15.83) | (14.80) | |
| Farmer Primary Sch. | 0.63 | 0.60 | 0.70 | |
| | (0.48) | (0.49) | (0.46) | |
| Farmer Secondary Sch. | 0.46 | 0.45 | 0.48 | |
| | (0.50) | (0.50) | (0.51) | |
| HH Size | 7.35 | 7.02 | 6.85 | |
| | (3.82) | (3.74) | (3.20) | |
| SC/ST | 0.15 | 0.15 | 0.10 | |
| | (0.36) | (0.36) | (0.31) | |
| Acres Owned | 1.73 | 1.48 | 1.72 | |
| | (1.93) | (1.73) | (1.84) | |
| Joint Difference | | 0.04 | 0.43 | |
| Participants | 461 | 573 | 48 | |
| Participant Characteristics | Main Sample | | | Supplemental |
| Male | 0.82 | | | 0.91 |
| | (0.39) | | | (0.29) |
| Age | 48.09 | | | 47.78 |
| | (16.67) | | | (15.17) |
| Primary School | 0.62 | | | 0.60 |
| | (0.48) | | | (0.49) |
| Secondary School | 0.45 | | | 0.41 |
| | (0.50) | | | (0.49) |
| HH Size | 7.35 | | | 7.92 |
| | (3.82) | | | (3.94) |
| SC/ST | 0.15 | | | 0.12 |
| | (0.36) | | | (0.32) |
| Saved Seeds | 3.08 | | | 3.04 |
| | (6.37) | | | (6.57) |
| Pulse Program Participant | 0.58 | | | 0.35 |
| | (0.49) | | | (0.48) |
| Pulse Program Beneficial | 0.74 | | | 0.70 |
| | (0.44) | | | (0.46) |
| Prior NGO Beneficiary | 0.64 | | | 0.42 |
| | (0.48) | | | (0.49) |
| Joint Difference | | | | 0.00 |
| Participants | 461 | | | 687 |

Notes: Group averages with standard deviations in parentheses. Rows defined as in Table S1 with the addition of self-reported stock of saved pulse seeds for planting, self-reported participation in the pulses program, subjective belief about whether the pulse program was beneficial, and self-reported participation in prior NGO initiatives.

Table S3: Effect of Evaluation Salience among Supplemental Farmer Sample

| | Experimental | | | Non-Experimental | |
|---|---|---|---|---|---|
| | Seed Demand | | Stated Belief | Seed Demand | Stated Belief |
| | (1) | (2) | (3) | (4) | (5) |
| Treated | -0.122 | -0.163 | 0.176 | -0.457 | 0.045 |
| | (0.10) | (0.16) | (0.10) | (0.18) | (0.12) |
| Salient × Treated | | 0.079 | -0.017 | 0.198 | 0.123 |
| | | (0.22) | (0.13) | (0.19) | (0.14) |
| | | [0.71] | [0.90] | [0.35] | [0.46] |
| Salient | -0.085 | -0.135 | -0.075 | -0.220 | -0.049 |
| | (0.09) | (0.17) | (0.12) | (0.17) | (0.11) |
| | [0.41] | [0.49] | [0.53] | [0.24] | [0.72] |
| Variable Mean | 1.23 | 1.23 | 0.70 | 1.11 | 0.64 |
| Salient Treat Effect (p-val.) | | 0.55 | 0.07 | 0.06 | 0.08 |
| R-Squared | 0.14 | 0.14 | 0.17 | 0.16 | 0.14 |
| Observations | 10455 | 10455 | 687 | 9965 | 641 |

Notes: Outcome is pulse seed quantity demanded. Sample is the set of farmers invited by village leaders to reach participation target in demand elicitation. Treated: Village received pulse program treatment in prior two years. Salient: High-salience script in elicitation. Pulse Program: Participant self-identifies as beneficiary of pulse program. Prior NGO: Participant self-identifies as beneficiary of previous NGO programs. Salient Treat Effect: p-value from test of sum of coefficients on Treated and Salient × Treated. All regressions include crop and price fixed effects, block (sub-district) fixed effects, and participant demographic controls. Columns (3) and (4) restrict to those that self-reported program participation during the red lentil (winter) elicitation. Standard errors clustered by village in parentheses; p-values from randomization inference over 1,000 re-draws of village-level salience treatment status in square brackets.

Table S4: Outcomes for Experimentally and Non-Experimentally Treated Farmers

| | Selection into Treatment | | |
|---|---|---|---|
| | Experimental | Non-Experimental | Difference |
| Male | 0.84 | 0.87 | 0.03 |
| | (0.37) | (0.34) | [0.47] |
| Age | 48.66 | 46.26 | −2.40 |
| | (17.05) | (15.61) | [0.11] |
| Primary School | 0.67 | 0.69 | 0.02 |
| | (0.47) | (0.47) | [0.79] |
| Secondary School | 0.52 | 0.51 | −0.01 |
| | (0.50) | (0.50) | [0.86] |
| HH Size | 7.02 | 8.32 | 1.30 |
| | (3.55) | (4.30) | [0.00] |
| SC/ST | 0.14 | 0.05 | −0.09 |
| | (0.35) | (0.22) | [0.00] |
| Acres Owned | 1.82 | 1.85 | 0.03 |
| | (1.92) | (1.72) | [0.89] |
| Joint Significance | | | 0.74 |
| Participation Rate | 0.67 | 0.68 | 0.01 |
| Participants | 454 | 191 | |
| Villages | 94 | 66 | |

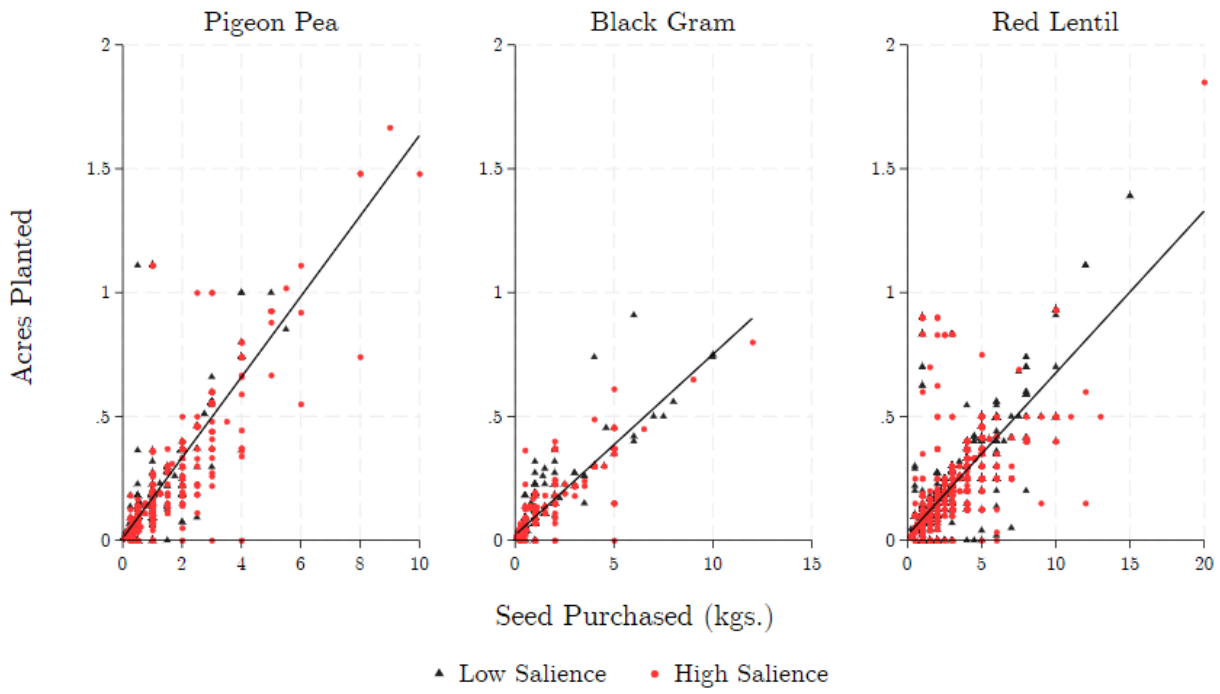
Notes: Group averages with standard deviations in parentheses and p-value of difference in square brackets. Rows correspond to fraction male, participant age, primary school completion, secondary school completion, household size, and fraction belonging to a schedule caste or scheduled tribe. Participation rate reflects fraction of those invited who appeared at either demand elicitation.

Table S5: Outcomes for Experimentally and Non-Experimentally Treated Farmers

| | Seed Demand (1) | Stated Belief (2) |
|---|---|---|
| Non-Experimental | 0.041 | -0.222 |
| | (0.15) | (0.09) |
| Salient × Non-Experimental | -0.172 | 0.222 |
| | (0.18) | (0.13) |
| Salient | 0.004 | -0.096 |
| | (0.11) | (0.07) |
| Variable Mean | 1.00 | 0.73 |
| R-Squared | 0.17 | 0.10 |
| Observations | 6690 | 409 |

Notes: Regressions restrict to study farmers in treated villages. Outcomes are input quantity demanded (Column 1) and self-reported belief that the program was beneficial. (Column 2). Non-Experimental: Village selected for treatment by implementer outside of experimental randomization. Salient: High-salience script in elicitation. All regressions include block (sub-district) fixed effects and participant demographic controls, and Column 1 includes crop and price-level fixed effects. Standard errors clustered by village in parentheses.

Figure S1: Seed Purchases and Cultivated Area (Treated Villages)




Notes: Data come from implementer administrative records in the two seasons of experimental demand elicitation, but include additional farmers that did not participate in elicitation. Records were only kept in villages formerly treated in the pilot experiment. Hollow triangles represent villages with low evaluation salience at demand elicitation, and solid squares represent high evaluation salience.

# B Supplemental Results

## B.1 Treatment Effect Heterogeneity

In Figures S2 and S3 we explore heterogeneity in treatment effects by crop and by implementing NGO. Figure S2 confirms the demand patterns observable in Figure 1: The effect of evaluation salience is strongest for black gram and red lentil. These are the two crops the pulses program focused most on in the second implementation year. By contrast, pigeon pea was included in initial program activities but subsequently de-emphasized by implementing partners. The effect of evaluation salience is correspondingly weaker for this variety.

We also observe heterogeneity in the effect of evaluation salience between implementing partners. The sign of the effect is the same for all four partners, as shown in Figure S3, which identifies the implementers by the district(s) in which they operate. However, the magnitude is greater for two—those in East and West Champaran districts—and smaller for the other two—those in Samastipur and Saran/Siwan districts. We observe no systematic differences in participants' self-reported participation in the pulses program, subjective belief that the program was beneficial, or self-reported participation in prior NGO initiatives that correspond to this pattern of treatment effect. With only four implementers in the study, we are hesitant to speculate on possible causes of heterogeneity and leave it to future research.

[Figure S2 about here.]

[Figure S3 about here.]

## B.2 Prior Engagement with Implementing NGO

Self-reported engagement with the local implementer is also related to seed demand, but this channel neither explains nor dilutes the relationship between treatment status and evaluation salience. After the second demand elicitation, farmers were asked about their participation in the pulses program and their in other NGO development activities in the past.

It is worth noting that self-reported beneficiary status may be an endogenous outcome of evaluation salience. Table S6 reports how the three variables relating to program benefits differ with experimental assignment. The first two columns show that while those treated in the pilot experiment are more likely to self-report participation the pulses program, a majority claim to have participated even in villages assigned to control. This is likely due to uncertainty about whether the question was asking about receiving the pulses support package or mere participating in the evaluation. The next two columns show that those who received the pulses intervention are also more likely to self-report having participated

in other NGO activities in the past despite random assignment of treatment. Interestingly, evaluation salience seems to prime participants to recall any NGO engagement, raising the likelihood of self-reporting participation in both the pulses program and in past NGO activities.

[Table S6 about here.]

The first two columns of Table S7 replicate those of Table 2, and columns (3) and (4) explore how experimental results vary with self-reported NGO engagement. Evaluation salience differentially affects the input demand choices of those who self-report prior NGO engagement. As Column 4 of Table S7 demonstrates, self-identified pulse program participants have sightly lower demand on average, but their demand substantially increases when the evaluation is made salient. The inverse is true of prior engagement with other NGO actives: seed demand is slightly greater among past beneficiaries, but evaluation salience substantially lowers it. This latter pattern is consistent with a model in which mentioning the NGO at the start of the evaluation leads participants to anchor expectations around the organization's typical strategy of providing benefits that are heavily subsidized or free, though other explanations are also possible. In any case, the coefficient estimates on treatment assignment, salience, and their interaction remain equally strong after controlling for self-reported measures of NGO engagement.

[Table S7 about here.]

## B.3 Effects on Input Demand Elasticity

Price anchoring may affect price elasticity in addition to quantity demanded. In Table S8 we report estimated treatment effects on the own-price elasticity of seed demand. The outcome is defined at the farmer–crop level by running a regression of log quantity demanded on log price separately for each farmer and each crop. There are no statistically significant effects of either pilot treatment or evaluation salience on demand elasticity.

[Table S8 about here.]

Table S6: Effect of Evaluation Salience on Self-Reported NGO Participation

| | Pulse Program Participation | | Prior NGO Participation | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Treated | 0.113 | 0.103 | 0.142 | 0.272 |
| | (0.06) | (0.09) | (0.07) | (0.10) |
| Salient × Treated | | 0.014 | | -0.255 |
| | | (0.12) | | (0.14) |
| | | [0.92] | | [0.11] |
| Salient | | 0.077 | | 0.216 |
| | | (0.10) | | (0.11) |
| | | [0.49] | | [0.09] |
| Variable Mean | 0.58 | 0.58 | 0.64 | 0.64 |
| Salient Treat Effect (p-val.) | | 0.16 | | 0.85 |
| R-Squared | 0.45 | 0.45 | 0.20 | 0.22 |
| Observations | 451 | 451 | 451 | 451 |

Notes: Outcomes are self-reported status in NGO programs. Columns (1) and (2) report participation in pulses program; (3) and (4) report participation in prior NGO initiatives. Treated: Village received pulse program treatment in prior two years. Salient: High-salience script in elicitation. Salient Treat Effect: p-value from test of sum of coefficients on Treated and Salient × Treated. All regressions include block (sub-district) fixed effects and participant demographic controls. Standard errors clustered by village in parentheses; p-values from randomization inference over 1,000 re-draws of village-level salience treatment status in square brackets.

Table S7: Effect of Evaluation Salience on Seed Quantity Demanded

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Treated | -0.260 | -0.531 | -0.163 | -0.466 |
| | (0.12) | (0.19) | (0.13) | (0.21) |
| Salient × Treated | | 0.506 | | 0.507 |
| | | (0.23) | | (0.26) |
| | | [0.04] | | [0.11] |
| Salient | -0.171 | -0.491 | -0.223 | -0.591 |
| | (0.10) | (0.19) | (0.12) | (0.23) |
| | [0.13] | [0.01] | [0.11] | [0.04] |
| Salient × Pulse Program | | | | 0.701 |
| | | | | (0.22) |
| | | | | [0.02] |
| Pulse Program Participant | | | 0.110 | -0.221 |
| | | | (0.15) | (0.21) |
| Salient × Prior NGO | | | | -0.563 |
| | | | | (0.21) |
| | | | | [0.04] |
| Prior NGO Beneficiary | | | 0.121 | 0.403 |
| | | | (0.14) | (0.18) |
| Variable Mean | 1.12 | 1.12 | 1.23 | 1.23 |
| Salient Treat Effect (p-val.) | | 0.85 | | 0.77 |
| R-Squared | 0.16 | 0.16 | 0.16 | 0.18 |
| Observations | 7390 | 7390 | 5065 | 5065 |

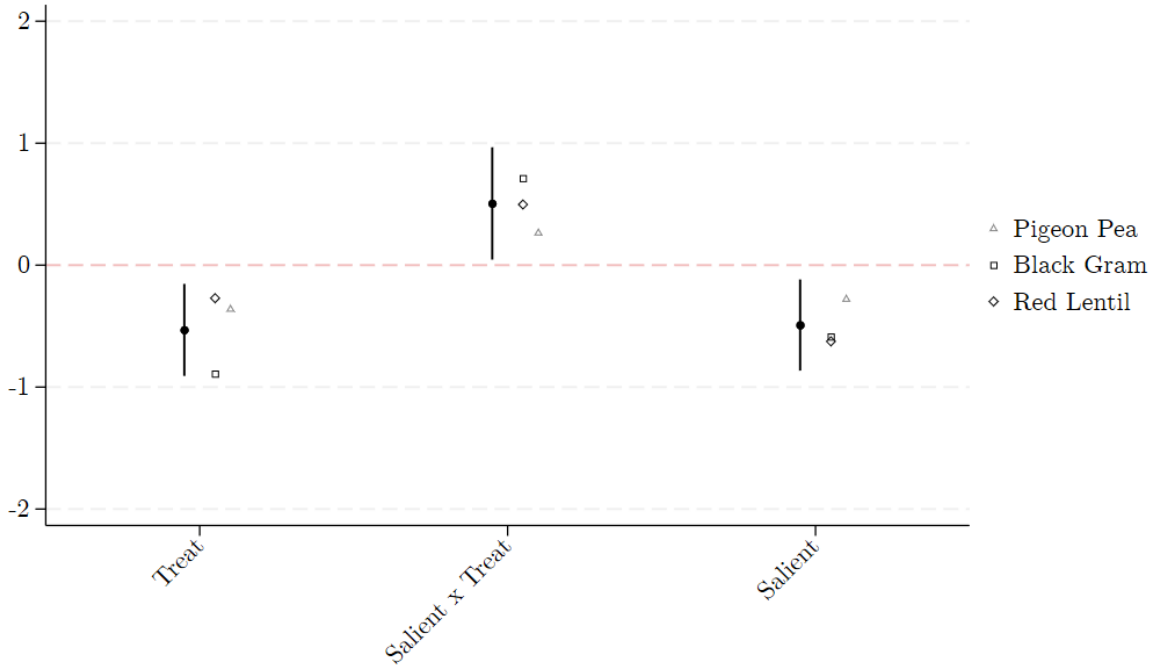
Notes: Outcome is pulse seed quantity demanded. Treated: Village received pulse program treatment in prior two years. Salient: High-salience script in elicitation. Pulse Program: Participant self-identifies as beneficiary of pulse program. Prior NGO: Participant self-identifies as beneficiary of previous NGO programs. Salient Treat Effect: p-value from test of sum of coefficients on Treated and Salient × Treated. All regressions include crop and price fixed effects, block (sub-district) fixed effects, and participant demographic controls. Columns (3) and (4) restrict to those that self-reported program participation during the red lentil (winter) elicitation. Standard errors clustered by village in parentheses; p-values from randomization inference over 1,000 re-draws of village-level salience treatment status in square brackets.

Table S8: Effect of Evaluation Salience on Input Demand Elasticity

| | (1) | (2) | (3) |
|---|---|---|---|
| Treated | -0.028 | -0.068 | -0.143 |
| | (0.08) | (0.12) | (0.12) |
| Salient × Treated | | 0.077 | 0.131 |
| | | (0.17) | (0.15) |
| | | [0.68] | [0.44] |
| Salient | -0.054 | -0.101 | -0.075 |
| | (0.08) | (0.12) | (0.13) |
| | [0.56] | [0.45] | [0.59] |
| Mean Elasticity | -0.81 | -0.81 | -0.83 |
| Salient Treat Effect (p-val.) | | 0.94 | 0.91 |
| R-Squared | 0.08 | 0.08 | 0.07 |
| Observations | 1159 | 1159 | 1459 |

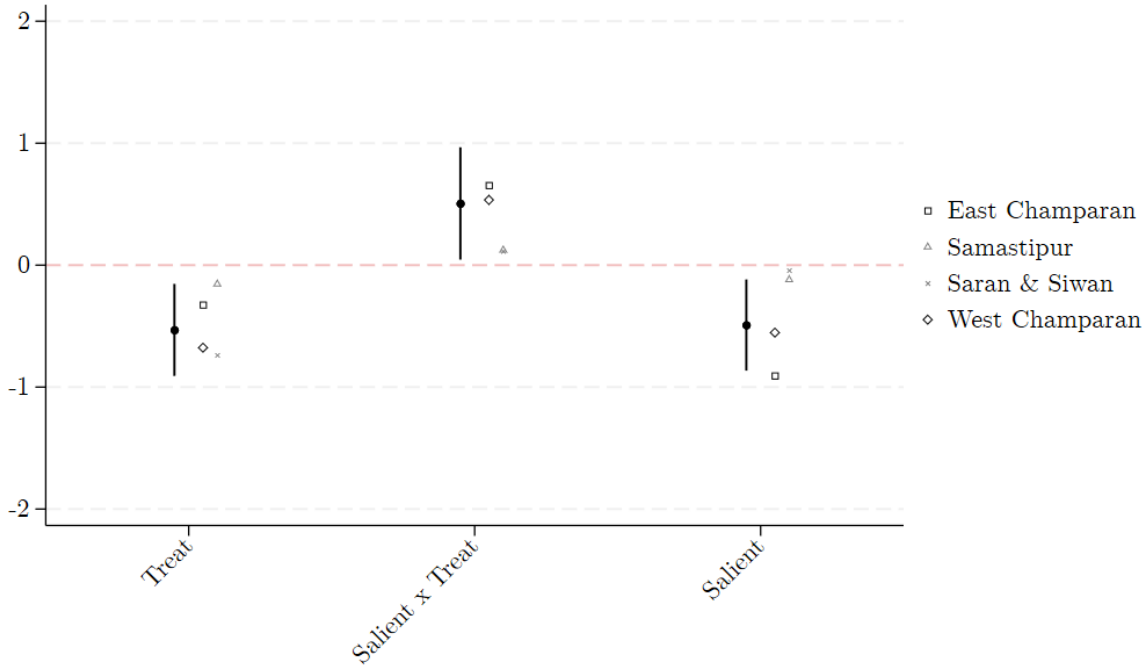
Notes: Outcome is own-price elasticity of pulse demand. Treated: Village received pulse program treatment in prior two years. Salient: High-salience script in elicitation. Pulse Program: Participant self-identifies as beneficiary of pulse program. Prior NGO: Participant self-identifies as beneficiary of previous NGO programs. Salient Treat Effect: p-value from test of sum of coefficients on Treated and Salient × Treated. All regressions include crop fixed effects, block (sub-district) fixed effects, and participant demographic controls. Columns (3) and (4) restrict to those that self-reported program participation during the red lentil (winter) elicitation. Standard errors clustered by village in parentheses; p-values from randomization inference over 1,000 re-draws of village-level salience treatment status in square brackets.

Figure S2: Treatment Effect Heterogeneity by Crop




Notes: Solid markers reproduce estimates from Column 2 of Table 2 following regression equation in (1). Hollow markers each present point estimates for a single crop.

Figure S3: Treatment Effect Heterogeneity by Implementing NGO




Notes: Solid markers reproduce estimates from Column 2 of Table 2 following regression equation in (1). Hollow markers each present point estimates for a single implementing NGO.