Implementer Identity Effects in Program Evaluation

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

[1]University of California, Davis

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

Abstract

Implementer effectiveness can be as important as policy design in shaping impacts of development interventions. Prior research has documented systematic differences in policy impacts based on the identity of the implementer (e.g. Vivalt, 2020). We examine how implementer identity affects program evaluation in the context of a package of agricultural subsidies and extension to promote the production of pulse crops in Bihar, India. This program was implemented as a two-year randomized controlled trial by local NGOs with a history of engagement in the area. Endline data includes a laboratory-style incentive-compatible elicitation of participants' demand for unsubsidized seeds, in which we experimentally vary the salience of the implementer. In one variation we explicitly advertise our evaluation of the implementer's efforts, while in the other we describe the exercise as a study on the viability of pulse farming. We find that increasing implementer salience lowers demand for pulse seeds, likely because the program was generally viewed as a failure due to adverse weather. However, the negative salience effect is 3–4 times greater in control relative to treatment villages. This disparity is driven largely by program beneficiaries whose demand increases by 20–25% when the implementer is made salient. Salience also differentially lowers demand among those who had participated in prior NGO initiatives that delivered benefits for free, consistent with price anchoring. Our results conform to a model where program beneficiaries reciprocate by delivering positive evaluations of implementers. We demonstrate they may take costly actions to do so in an incentive-compatible demand elicitation, and the effect is quantitatively large: the estimated negative treatment effect is attenuated by 66% when evaluation is made salient. These findings suggest that program evaluation with popular implementers may be subject to a type of Hawthorne effect that systematically biases in favor of success.

JEL Codes: O22, C90, L31

1 Introduction

Implementer effectiveness can be as important as policy design in economic development. Prior research has documented systematic differences in measured policy impacts based on the identity of the implementer (e.g. Vivalt, 2020). In this paper, we provide evidence that implementer identity can generate systematic differences not just in policy implementation, but in program evaluation as well.

We study the role of implementer identity during endline data collection following a two-year intervention to promote pulse production in the state of Bihar in India. The policy revolved around input subsidies and intensive agricultural extension delivered by four non-governmental organizations (NGOs) with an extensive history of local rural development work. Evaluation data includes a revealed-preference elicitation of demand for agricultural inputs after subsidies expired. We evaluate whether varying the salience of the evaluation during data collection alters the estimated treatment effect.

To address this question, we experimentally vary the salience of evaluation during an elicitation of agricultural input demand. We introduce the elicitation as either an explicit evaluation of the implementer's efforts or as a general study of regional input demand, and investigate whether this manipulation alters measured input demand. By tying the experimental variation to a real-stakes elicitation exercise, we effectively test whether evaluation itself induces participants to take costly actions that affect their agricultural input and production decisions.

To accurately interpret results, we must ensure that the experimental manipulation does not communicate additional information beyond making salient the impact evaluation. We achieve this by eliciting demand for a consistent product explicitly sourced by the local implementer. In doing so, we ensure that invoking the implementer's identity does not inadvertently serve as a form of quality verification or provide other details about the input being sold. As a result, we interpret our findings as the effect of evaluation salience alone on the demand for a uniform agricultural product.

This study is most directly related to research on the Hawthorne effect in experimental economics. Past work has established that subjects in an experiment may alter their behavior when they know they are being monitored or evaluated.[1] We extend this line of work to investigate a case where participants are made aware that their decisions can affect the evaluation of a third party with whom they may have a prior relationship.

Our findings more generally informative for extracting policy lessons from small-scale pilots. Past research has demonstrated that evaluation results may not replicate at scale for many reasons, including favorable site selection in the initial stages (Allcott, 2015), bias in reporting of results (DellaVigna and Linos, 2020), and behavioral responses among the treated population (Dhaliwal and Hanna, 2017). We add a new reason to be wary of pilot results: beneficiary populations may alter their behavior to deliver favorable

[1] See, e.g., Campbell et al. (1995); de Amici et al. (2000); Feil et al. (2002); Mangione-Smith et al. (2002); Verstappen et al. (2004); Eckmanns et al. (2006); Leonard and Masatu (2006); McCarney et al. (2007); Leonard (2008); Leonard and Masatu (2008); Kohli et al. (2009); Clasen et al. (2012); Fernald et al. (2012); Srigley et al. (2014).

evaluations when they have connections to the implementing organization.

This paper proceeds as follows. In Section 2 we discuss the context in which this study takes place. Section 3 discusses the data and methodology, and Section 4 presents results. Finally, we discuss implications in Section 5.

2 Background

This research takes place in the Indian state of Bihar. The state is among the poorest in the country, where over a third of households fall 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. There are currently 4,255 registered NGOs and volunteer organizations in the state.[2] The majority 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 at larger scale.

Our study is tied to a joint development initiative between the Government of India and local organizations in Bihar. In 2016, the Government of India partnered with four local NGOs to develop a policy to increase the production of pulses by farm households in the state. The partnership designed a package of input subsidies and agricultural extension services to boost pulse cultivation over the short term. This package was piloted in five districts to test whether intensive short-term support could shift the long-term crop portfolio of participating households.

The pulses support policy was administered by the four local partner organizations over two years. From summer 2017 through spring 2019, implementing organizations were responsible for procuring and distributing pulse seeds at subsidized prices, conducting local training and extensions sessions to demonstrate best practices, giving individualized feedback to program participants through the cropping season, and assisting with the sale of output. These 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.[3]

Implementing partners for this project were selected because of their track record with rural development. All four organizations 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 finance. As a result, local implementers had preexisting relationships with study participants before the inception of the pulses program. We quantify how these relationships affect impact evaluation in the pilot experiment.

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

[3]The pilot experiment was pre-registered with the AEA RCT Registry: AEARCTR-0003872.

3 Data and Methodology

3.1 Data

The primary data for this study come from a laboratory-style elicitation of willingness-to-pay for pulse seeds among farmers participating in the pilot experiment. A key evaluation outcome for the trial was sustained production of pulses after main program activities had concluded. As part of this evaluation, we elicited demand for certified seeds ahead of the planting season among a sample of study farmers. Elicitations measure seed demand for pigeon pea and black gram in the summer (Kharif) season and for green lentil in the winter (Rabi) season of 2019.

Each demand elicitation was conducted with a sample of 8–10 farmers from the same village. Participants took part in a Becker-DeGroot-Marshcak auction as the basis for their certified seed orders. For each seed variety, participants were given list of possible prices and asked to reveal their quantity demanded at each price. At the end of the session, one price at random was selected from the list, and participants had the option to purchase their stated demand at that price. The purchase transaction ensures that each demand decision potentially had real financial stakes.

The maximum price in each elicitation was the prevailing market price. Above this level, demand would rationally have been zero because participants always have the option to purchase seeds in the market. Hence, the elicitation reflects demand when purchasing seeds at a discount relative to the market and we do not observe hypothetical demand above market prices.

To ensure incentive compatibility during the elicitation, participants did not have the option to adjust their quantity demanded after the price was revealed. However, there as a delay of several days between the price elicitation and actual delivery of seeds. At the time of delivery, participants had the option to either purchase a different quantity at the market price (foregoing the experimental discount) or to purchase no seeds at all. To our knowledge, there were a few isolated instances where demand fell to zero, and no instances of farmers purchasing a different amount at the market price. Therefore, we interpret experimental responses as farmers' true intended demand at the time of the experiment.

After the demand elicitation exercise, participants are asked a short set of questions related to demographics and history of pulse cultivation. Participants in the winter season are also asked about their involvement with the pulses pilot program and their past involvement with the local implementer in general.

Participants for this study were recruited from the sample of farmers involved in the pilot evaluation. This sampling frame was selected prior to treatment assignment in order to be experimentally comparable across treatment arms. When too few farmers from this sampling frame were available to participate, additional participants were recruited via snowball sampling from among the contacts of existing participants on the day of the demand elicitation.

3.2 Study Design

This study leverages two sources of experimental variation. The primary variation comes from the script of the demand elicitation, where we experimentally vary the salience of the implementing partner and evaluation. In half the sessions, we explicitly motivate our work as an exercise to evaluate the efforts of the implementing NGO to promote pulses over the prior two years. In the other half, we more generally explain we are conducting a study on the desirability of pulses in the region. Other than this discrepancy, the demand elicitation script is identical.[4]

Table 1 presents descriptive statistics on participants across the two salience arms. The first five rows report demographic information of study participants. The next two rows describe whether participants self-report as being beneficiaries of the pulses program and of other NGO initiatives in the past, respectively. The next three rows provide self-reported measures of education, and the final row measures the potion of the sample that was recruited through snowball sampling. Note that the salience variation took place after participant recruitment, so experimental comparisons do not rely on any assumptions about consistency or representativeness in the snowball portion of the recruitment procedure. Only the fraction of scheduled cast/scheduled tribe members differs significantly at the 5% level, and a joint F-test rejects that treatment groups are systematically unbalanced at the 5% level.

[Table 1 about here.]

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

$$Q_{icp} = \beta Salient_i + \sigma_c + \phi_p + \alpha_{d(i)} + \xi_{z(i)} + \epsilon_{icp} \tag{1}$$

where Q_icp denotes the quantity demanded by individual i for seeds of crop c at price p. The coefficient of interest, β indicates how this demand differs on average for individuals exposed to the high-salience script; this assignment remains constant over crops. The coefficients σ_c and ϕ_p control for crop-specific and price-level demand shifters, respectively. $\alpha_{d(i)}$ control for district-level demand shifters, and $\xi_{z(i)}$ control for demand shifters induced by treatment status in the original pilot experiment.

To interpret β as the effect of the salience of evaluation, we must ensure that mentioning the implementing partner does not communicate other information about the demand elicitation. In particular, it is possible that associating the seeds on offer with a local NGO may alter participants' beliefs about the quality of those seeds if participants believe the NGO to be of different quality than a typical market actor (or research team).

We address this concern by distributing coupons for seed delivery rather than the seeds themselves. In each demand elicitation, we make clear that participants will receive a coupon for seeds sourced by the local implementing partner. In this way, we ensure uniformity in the product that study participants believe they are expressing demand for, so that treatment effects can be attributed to evaluation salience.

[4]This experiment was pre-registered with the AEA RCT Registry: AEARCTR-0004405.

The salience treatment cross-cuts the second source of experimental variation inherited from treatment assignment in the pilot evaluation. This study includes farmers from 99 experimentally treated villages that received two years of input subsidies and extension support as well as from 59 control villages that did not. We also elicit demand in 70 non-experimentally treated villages that were selected by the implementing partners before pilot randomization.

We test whether the salience of the evaluation differentially affects estimated treatment effects by extending specification (1) to include

$$Q_{icp} = \beta Salient_i + \gamma_Z \mathbf{1}\{z(i) = Z\} \times Salient_i + \sigma_c + \phi_p + \alpha_{d(i)} + \xi_{z(i)} + \epsilon_{icp} \tag{2}$$

where γ_Z measures the differential effect of evaluation salience on study participants assigned to treatment group Z in the pilot experiment. A finding of $\gamma_Z \neq 0$ would indicate that the estimated treatment effect in the pilot experiment would differ based on whether the evaluation were made salient or not at the time of data collection.

We also investigate how the impact of salience on willingness-to-pay differs by engagement with implementing partners. To address this question, we interact the salience treatment with (i) self-reported engagement in the pulses program and (ii) self-reported involvement in prior NGO efforts. Formally, this amounts to replacing individuals' experimental treatment status $z(i)$ with their non-experimental level of past NGO engagement $x(i)$ in regression equation(2).

4 Results

Experimental results are presented in Table 2. Standard errors are clustered at the participant level in all regressions.

[Table 2 about here.]

The first column corresponds to regression equation (1). Results indicate that both assignment to treatment in the pilot experiment and increased salience of the evaluation lowered demand for pulse seeds, consistent with the broader program evaluation in which pulses displaced more profitable crops during the experimental period. The results in this study suggest that knowledge of this negative impact was widespread, and bringing it to mind alters input purchases for the next season. The effects are quantitatively large, with evaluation salience depressing seed demand by nearly 10% relative to control.

The second column breaks down the effect of evaluation salience by pilot treatment status as described in regression equation (2). While the net effect of evaluation salience remains uniformly negative, it is substantially tempered among participants in villages receiving program benefits in the pilot experiment. The negative impact of evaluation salience on demand is 70–75% smaller among farmers in treated villages than in control villages.

Interestingly, the mitigating effect of pilot treatment does not vary between experimentally and non-experimentally treated villages. One potential reason implementing partners may have non-experimentally selected some villages to treat is that they anticipated seeing strong program effects in these villages. Our results do not provide evidence of this motivation as evaluation results do not significantly differ between the two treated groups. However, it should be noted that all pilot villages are located in areas where implementing partners have substantial prior engagement, suggesting that the scope for site selection may have been limited in practice.

Columns (3) and (4) of Table 2 show how NGO engagement influences experimental results.[5] Participants from treated villages in the demand elicitation are 20% more likely to report being beneficiaries of the pilot program than those in control villages. Surprisingly, they are also 13% more likely to remember other engagement in NGO activities despite random assignment to treatment. As column (3) demonstrates, after controlling for participants' self-reported involvement with implementers, treatment status in the pilot experiment has little additional explanatory power. Thus, it seems the treatment effect on willingness to pay for seeds operates primarily through participants' perceptions of NGO engagement.

Column (4) interacts these covariates with the salience treatment according to the regression in (2). Among those who self-report as pilot program beneficiaries, increasing the salience of evaluation once again has a positive and significant effect. Here the point estimate is large enough to reverse the main negative effect of salience. On net, those who self-report as having benefited from the pilot program actually increase their demand for pulse seeds when the evaluation is made salient during data collection.

By contrast, engagement in other activities by an implementing NGO has the opposite effect. Participants who self-report as having benefited form NGO activities other than the pulses program demonstrate lower demand for pulse seeds on average. This effect is similarly amplified by increasing the salience of evaluation, with a nearly 25% decrease in quantity demanded in the high-salience group relative to the control mean.

5 Discussion

Overall, our results indicate that the salience of an impact evaluation can influence the estimated treatment effect. We establish this fact through an experimental demand elicitation with real stakes where participants make binding input decisions for the coming crop season. Our experimental design eliminates differences in the perceived quality of inputs to isolate the effect of evaluation salience on product demand.

Experimental variation in the salience of evaluation takes place in the context of an agricultural policy that was largely negatively perceived. While increasing the salience of evaluation heightens negative responses on average, this effect is substantially attenuated among participants in villages that directly benefited from the intervention. Among self-reported program beneficiaries, the salience effect is large enough to reverse the negative result, leading to a net increase in demand for this subpopulation.

[5]There are fewer observations for these regressions because questions on NGO engagement were only asked during the winter season while eliciting demand for green lentils.

These findings are consistent with a model in which beneficiaries of development programs reciprocate by altering their behavior in ways that favor positive evaluations of the program. Motivation for this form of reciprocity may be either backward-looking or forward-looking. That is, participants may seek to reward the implementer's past efforts with a positive evaluation, or they may anticipate that a positive evaluation will increase the likelihood of receiving continued benefits in the future. We leave the exact motivation for reciprocity as an open avenue for future research.

This study also uncovers suggestive evidence of price anchoring in the delivery of services. We measure demand for agricultural inputs following a two-year intervention that initially transferred inputs for free, implemented by local NGOs with a history of providing other free or heavily subsidized services. We find that increasing the salience of the implementer in this setting lowers input demand at market prices among those who self-identify as having benefited from past development initiatives. This effect is consistent with anchoring around the expectation of subsidies. In this case, anchoring may offset some of the demand inflation caused by the desire for reciprocity.

Our results show that implementer identity plays a role in program evaluation. As policies expand from promising pilots to large-scale implementation, responsibility for implementation usually shifts to new administrative units. We demonstrate that this type of shift can alter not only the way a policy is managed and implemented, but also how it is evaluated and received by its beneficiary communities.

References

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

Campbell, J. Peter, Vada A Maxey, and William A. Watson, "Hawthorne effect: implications for prehospital research.," *Annals of emergency medicine*, 1995, *26* (5), 590–4.

Clasen, Thomas, Douglas Fabini, Sophie Boisson, Jay Taneja, Joshua Song, Elisabeth Aichinger, Anthony Bui, Sean Dadashi, Wolf-Peter Schmidt, Zachary Burt, and Kara Nelson, "Making Sanitation Count: Developing and Testing a Device for Assessing Latrine Use in Low-Income Settings," *Environmental science & technology*, 02 2012, *46*, 3295–303.

de Amici, Donatella, Catherine Klersy, Felice Ramajoli, L Brustia, and P L Politi, "Impact of the Hawthorne effect in a longitudinal clinical study: the case of anesthesia.," *Controlled clinical trials*, 2000, *21* (2), 103–14.

DellaVigna, Stefano and Elizabeth Linos, "RCTs to Scale: Comprehensive Evidence from Two Nudge Units," Working Paper 27594, National Bureau of Economic Research July 2020.

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.

Eckmanns, Tim, Jan Bessert, Michael Behnke, Petra Gastmeier, and Henning Rüden, "Compliance With Antiseptic Hand Rub Use in Intensive Care Units The Hawthorne Effect," *Infection Control & Hospital Epidemiology*, 2006, *27*, 931–934.

Feil, Philip, Jennifer Sherah Grauer, Cynthia C Gadbury-Amyot, Katherine S. Kula, and Michael D McCunniff, "Intentional use of the Hawthorne effect to improve oral hygiene compliance in orthodontic patients.," *Journal of dental education*, 2002, *66* (10), 1129–35.

Fernald, Douglas, Letoynia Coombs, Lauren Dealleaume, David West, and Bennett Parnes, "An Assessment of the Hawthorne Effect in Practice-based Research," *Journal of the American Board of Family Medicine : JABFM*, 01 2012, *25*, 83–6.

Kohli, Erol, Judy Ptak, Randall Smith, Eileen Taylor, Elizabeth Talbot, and Kathryn Kirkland, "Variability in the Hawthorne Effect With Regard to Hand Hygiene Performance in High- and Low-Performing Inpatient Care Units," *Infection control and hospital epidemiology : the official journal of the Society of Hospital Epidemiologists of America*, 02 2009, *30*, 222–5.

Leonard, Kenneth, "Is Patient Satisfaction Sensitive to Changes in the Quality of Care? An Exploitation of the Hawthorne Effect," *Journal of health economics*, 04 2008, *27*, 444–59.

__ **and Melkiory Masatu**, "Outpatient process quality evaluation and the Hawthorne Effect," *Social science & medicine (1982)*, 12 2006, *63*, 2330–40.

__ **and** __ , "Using the Hawthorne Effect to Examine the Gap between a Doctor's Best Possible Practice and Actual Performance," *SSRN Electronic Journal*, 01 2008.

Mangione-Smith, Rita, Marc Elliott, Laurie Mcdonald, and Elizabeth McGlynn, "An Observational Study of Antibiotic Prescribing Behavior and the Hawthorne Effect," *Health services research*, 12 2002, *37*, 1603–23.

McCarney, Rob, James Warner, Steve Iliffe, Robbert Haselen, Mark Griffin, and Peter Fisher, "The Hawthorne Effect: A Randomised, Controlled Trial," *BMC medical research methodology*, 02 2007, *7*, 30.

Srigley, Jocelyn, Colin Furness, G. Baker, and Michael Gardam, "Quantification of the Hawthorne effect in hand hygiene compliance monitoring using an electronic monitoring system: a retrospective cohort study," *BMJ quality & safety*, 07 2014, *23*.

Verstappen, Wim, Trudy van der weijden, Gerben Riet, Jeremy Grimshaw, Ron Winkens, and Richard Grol, "Block design allowed for control of the Hawthorne effect in a randomized controlled trial of test ordering," *Journal of clinical epidemiology*, 12 2004, *57*, 1119–23.

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
Male	0.94	0.91
	(0.23)	(0.28)
Age	49.99	50.52
	(15.57)	(15.05)
Married	0.91	0.93
	(0.29)	(0.25)
HH Size	8.90	9.00
	(4.13)	(4.35)
SC/ST	0.75	0.81
	(0.43)	(0.39)
Beneficiary	0.46	0.46
	(0.50)	(0.48)
Other NGO	0.53	0.52
	(0.50)	(0.50)
Literate	0.73	0.70
	(0.45)	(0.46)
Primary	0.77	0.74
School	(0.42)	(0.44)
Secondary	0.60	0.56
School	(0.49)	(0.50)
Snowball	0.61	0.60
	(0.49)	(0.49)
Households	1,097	1,182
Villages	106	122

Group averages with standard deviations in parentheses. Rows correspond to fraction male, participant age, fraction married, fraction belonging to a schedule caste or scheduled tribe, fraction self-reporting as beneficiary of pulses pilot, fraction self-reporting as beneficiary of other NGO programs, self-reported literacy, primary school completion, secondary school completion, and fraction recruited via snowball sampling.

Table 2: Effect of Salience on Quantity Demanded

Outcome: Seed Demand				
	(1)	(2)	(3)	(4)
Salience	-0.12	-0.26		-0.14
	(0.04)	(0.08)		(0.11)
Salience×		0.20		
Treated		(0.10)		
Salience×		0.18		
Non-Experimental		(0.10)		
Salience×				0.41
Beneficiary				(0.17)
Salience×				-0.50
Other NGO				(0.17)
Treated	-0.17	-0.27	-0.04	-0.05
	(0.05)	(0.08)	(0.10)	(0.10)
Non-Experimental	-0.25	-0.35	-0.11	-0.11
	(0.05)	(0.08)	(0.10)	(0.10)
Beneficiary			0.28	0.10
			(0.10)	(0.14)
Other NGO			-0.16	0.09
			(0.09)	(0.14)
District FEs	X	X	X	X
Crop FEs	X	X		
Price FEs	X	X	X	X
Control Mean	1.44	1.44	2.05	2.05
R-Squared	0.14	0.14	0.11	0.11
Observations	26,335	26,335	8,260	8,260

Standard errors clustered by participant in parentheses. Treated: Experimentally assigned to pilot treatment. Non-Experimental: Non-experimentally selected for treatment in pilot by NGO. Beneficiary: self-reported beneficiary of pulses pilot intervention. Other NGO: self-reported involvement with other NGO programs in past. Columns (3) and (4) have only data from green lentil elicitation in winter 2019.