Does Survey Mode Matter? Comparing In-Person and Phone Agricultural Surveys in India

Ellen Anderson[1], Travis J. Lybbert[1], Ashish Shenoy[1], Rupika Singh[2], and Daniel Stein[2]

[1]University of California, Davis
[2]IDinsight

February 2023

Abstract

Ubiquitous mobile phone ownership makes phone surveying an attractive method of low-cost data collection. We explore differences between in-person and phone survey measures of agricultural production collected for an impact evaluation in India. Phone responses have greater mean and variance, and this difference persists within a subset of respondents that answered the same question over both modes. Treatment effect estimation nonetheless remains stable across survey mode, but estimates are less precise when using phone data. These patterns are informative for cost and sample size considerations in study design and for aggregating evidence across study sites or time periods.

JEL codes: C83, O13, Q12

1 Introduction

Household surveys are standard in economics research, especially in developing economies where where administrative records and official statistics are likely to be incomplete due to high degrees of informality (see Deaton, 2005). Traditional survey methods rely on face-to-face interviews with study participants, but the worldwide penetration of information and communication technology makes remote data collection increasingly accessible. In particular, commoditization of mobile phones—an estimated 73% of adults globally and 58-61% in developing countries now own a mobile phone (ITU, 2022)—enables connectivity in even the most isolated parts of the world.

In this paper, we investigate differences between in-person and phone survey data collected during an agricultural extension experiment in Bihar, India. Phone surveying presents an appealing alternative to in-person data collection because of potential cost savings. Interviewing study participants by phone mitigates the logistical difficulty of physically locating a desired respondent and minimizes enumerator transportation and lodging. However, phone contact can introduce new forms of attrition, and respondents may behave differently when not physically present with an interlocutor. Therefore, it is valuable to explore precisely how to interpret phone responses in relation to comparable in-person data.

Our study leverages data from two parallel impact evaluations of the same underlying program. Evaluators asked a harmonized set of questions on agricultural production, with one team going door-to-door and the other calling by phone, and 42% of households participated in both surveys. We analyze a combined database of responses to the same questions asked to households sampled from the same population that vary only by the mode in which the respondent was contacted.

We conduct two types of comparisons between survey modes. First, we quantify differences in the distribution of self-reported agricultural production for the four most common crop varieties. Phone respondents report 8–51% more production on average, depending on the crop, and there is

greater variance among phone responses for three out of four crops. This pattern is consistent across the output distribution, with larger fractions of phone respondents reporting positive production for three out of four crops and higher production values over the phone at the median, 75^{th}, and 90^{th}percentiles for all four crops.

These differences persist even after accounting for selective attrition by survey mode. Phone respondents in our study tend to be wealthier and more educated on average, mirroring general patterns of mobile phone ownership and use. Nevertheless, the gap between modes at each production decile remains nearly as large when restricting to the subset of households or respondents that participated in both surveys. Within-household and even within-person differences in self-reported production by survey mode explain more than sixty percent of the total measured gap for three out of four crops.

There is little evidence that production values were influenced by differential engagement among phone respondents. 87% of participants rounded their response to the nearest five kilograms, and 69% to the nearest ten, but these fractions are nearly identical across survey modes. Therefore, we reject that differences in self-reported production were induced by respondents more carelessly rounding small quantities up over the phone or down in person. We also rule out any systematic bias caused by differences in survey timing.

Second, we compare experimental treatment effects estimated using each method of data collection. In contrast to our prior findings, the within-sample relationship between treatment status and self-reported production remains stable across survey modes. Regression coefficients are similar in magnitude, and we fail to reject equality for any major crop variety. However, we report greater estimation error when using the phone survey data, consistent with higher variance in phone responses.

Taken together, these results can inform research design and evidence aggregation. We show

that heterogeneity in the method of contact may introduce bias into comparisons of survey outcomes across populations. Such bias can undermine conclusions about differences between study populations or about the evolution of outcomes within a population over time, such as in subsequent rounds of a panel or repeated cross-sectional survey. To make such comparisons viable, it is necessary to establish reliable indicators that link data across survey modes. We find this issue to be less of a concern for program evaluation.

Our findings also highlight a tradeoff in the use of phone surveying for program evaluation. While it may be cheaper to conduct surveys by phone than in person, the resulting data may be noisier. In such cases, phone-based data collection necessitates larger samples to achieve the same power, offsetting some of the cost savings. In our context there is substantial heterogeneity in the breakeven point: depending on the crop, the phone sample would have needed to be 1.3–12.3 times larger than the in-person sample to estimate treatment effects with the same precision. In general, it would be prudent for researchers to consider noise introduced by survey method when calculating power.

Evidence on how survey mode affects data reliability most commonly focuses on self-reported health indicators. Investigators report mixed results on the correspondence between in-person and phone responses, and those showing statistical differences draw no systematic conclusions about types of indicators subject to mode effects or direction of bias (Greenfield et al., 2000; Biemer, 2001; Scherpenzeel and Eichenberger, 2001; St-Pierre and Béland, 2004; Nord and Hopwood, 2007; Ferreira et al., 2011; Mahfoud et al., 2015; Greenleaf et al., 2020). Other comparisons include phone-based measures of consumer valuation (Maguire, 2009; Szolnoki and Hoffmann, 2013), microenterprise data (Garlick et al., 2020), and school performance (Crawfurd et al., 2021). In developing-country agriculture, Kilic et al. (2021) uncover a similar pattern to ours of greater self-reported

production by phone than in-person among tuber farmers in Malawi.[1]

Our analysis extends this literature in three ways. First, the overlapping sample of respondents allows for within-household estimation of survey mode effects. Only Mahfoud et al. (2015) include this feature, but prime for consistency by advertising phone contact as a check on prior in-person responses.[2] Second, while most existing work tests for bias in sample means, we also report differences in precision and at various production quantiles. In particular, our finding of greater variance in phone-based data, consistent with Garlick et al.'s study of microenterprises (2020), can inform sample size calculations in research design. Third, we investigate how the mode used for data collection affects program evaluation in agriculture. Crawfurd et al. (2021) reach a similar conclusion that survey mode affects measurement of student test scores on average, but does not bias evaluation of an educational intervention.

Research interpreting phone survey data is especially timely following COVID-19 disruptions that forced remote data collection. To accurately quantify the evolution of economic outcomes through the pandemic and beyond, researchers must find ways to relate outcomes across surveys (see Egger et al., 2021; Josephson et al., 2021, for successful examples). To the extent that lessons learned from the large-scale use of remote data collection during the pandemic (Gourlay et al., 2021) enable these practices to remain in place in the future, it will be important to develop methods to establish comparability between pre- and post-pandemic surveys.

Our investigation also relates to the growing body of work on how to aggregate evidence across studies. Many policy evaluations take place in idiosyncratic contexts, and organizations such as 3ie[3] and Cochrane Reviews[4] devote substantial resources to drawing general conclusions about policy

[1] A complementary application of mobile phone data avoids surveying altogether and draws inferences about household-level outcomes from metadata (see Blumenstock et al., 2015).

[2] Biemer (2001); Nord and Hopwood (2007) analyze panel data from national statistical offices where the first survey round is conducted in-person and subsequent rounds by phone, but this structure does not allow separate identification of survey mode and time effects within household.

[3] https://www.3ieimpact.org/evidence-hub/publications/systematic-reviews

[4] https://www.cochranelibrary.com/cdsr/reviews

impacts. Meager (2019) provides an empirical framework for evidence aggregation that disentangles average policy impacts, context-specific heterogeneity, and sampling variation; and Pritchett and Sandefur (2015) argue heterogeneity across contexts can threaten external validity moreso than poor identification. In this paper we demonstrate how and when the mode of survey can introduce study-specific heterogeneity in measured outcomes that is largely uninformative for policy decisions.

In Section 2 we describe the data and methodology used for analysis. Section 3 reports results on differences in the sample distribution of responses, and Section 4 on estimated treatment effects. Section 5 concludes.

2 Data and Methodology

Data for this study come from two overlapping experimental evaluations of an agricultural extension program to promote pulse cultivation in Bihar, India. The initial intervention began in May 2017, followed by an in-person midline household survey in December 2017. At midline, households answered questions about demographic characteristics as well as pre-harvest farm area devoted to pulses. The 2,346 midline survey respondents, selected at random from the 6,971 evaluation households, constitute the sampling frame for the current study. Of these, 1,100 were randomly selected for an extended midline survey that asked about socioeconomic status in greater detail.

In this paper we report results on household pulse production from first-year endline surveys conducted in May–June 2018. Endline data was collected by parallel in-person and phone surveys that asked nearly identical questions about household production by pulse variety conditional on having positive area planted at midline. The two data collection exercises were motivated by a desire to optimize for a number of different research questions. The phone survey allowed a larger sample size, with the hope that this would give higher power for the key outcome of pulse production. The in-person survey contained more modules, and allowed exploration of further outcomes.

We analyze production of the four most common varieties of pulses—pigeon peas (*arhar*), grown by 681 households; red lentils (*masoor*), grown by 854 households; green peas (*mattar*), grown by 398 households; and fava beans (*bakla*), grown by 390 households. Of these, pigeon peas and red lentils were explicitly targeted by extension efforts in the year of study. In total, 1,533 of the midline respondents reported positive area planted for at least one of the four major pulse varieties, and fewer than 100 households reported growing any other variety of pulses.

In-person surveying was part of a long-term impact evaluation by researchers at the University of California, Davis. Researchers attempted to reach all 1,100 extended midline survey respondents. 1,058 households answered the survey, corresponding to an in-person attrition rate of 3.8%. Those that had reported positive area devoted to pulses at midline were asked about their production by variety at endline, and in-person surveys included a number of other questions on agricultural production and food consumption. Full evaluation results from the in-person survey are reported by Lybbert et al. (2022).

Phone surveying was used for a short-term cost–benefit analysis by researchers at IDinsight. Researchers attempted to reach all 1,533 households in the sampling frame that had reported positive area devoted to pulses at midline. Of these, 1,266 responded corresponding to an attrition rate of 17.4% by phone. Phone respondents were asked only about pulses production due to time constraints imposed by the survey format. Full evaluation results from the phone survey are reported by Anderson et al. (2022).

To the extent possible, questions about pulse production were identical across surveys. The exact wording is provided in Appendix A. Enumerators in both surveys were instructed to speak to the primary farmer in the household, who had previously been identified in the midline survey. This individual was the respondent in 84% of in-person and 81% of phone surveys. We interpret differences in the difficulty of reaching the desired respondent to be an inherent feature of data

collection, and therefore treat it as one channel through which survey mode effects may operate.

While both surveys were administered in parallel, the same household was typically not contacted by both modes on the same day. On average, the in-person survey was conducted 7 days after the phone survey, but differences range from 13 days earlier to 26 days later. In Appendix A we verify responses are not systematically related to this variation in timing.

The upper tail of all production responses are Winsorized to the 95% level independently by crop and by mode to match how data would have been treated had either study been conducted in isolation.

This study presents two types of comparisons between in-person and phone survey responses. First, we compare moments in the distribution of self-reported production volume across survey mode. We report the mean, variance, and value at each decile for the four most common pulse varieties among households that reported positive area planted for that pulse at midline. This comparison reveals how inferences about population outcomes differ by survey mode inclusive of any bias introduced by differential attrition by survey respondents.

We next decompose differences in distribution into selection and mode effects. This analysis leverages the fact that 715 households were contacted for both in-person and phone surveying. Out of these, 584 responded to both modes of contact, and in 429 cases the exact same individual answered each time. Variation in self-reported production volume within this overlapping sample can be attributed purely to survey mode, and the characteristics of non-respondents provide evidence about differential attrition bias. We also explore respondent engagement using evidence of rounding to the nearest five or ten.

Second, we investigate how survey mode affects program evaluation. Here we estimate the intention-to-treat (ITT) effect on pulse production separately within each survey, represented by β

in

$$Y_i = \beta T_i + X_i \delta + \gamma_{b(i)} + \epsilon_i \tag{1}$$

where Y_i represents production for household i living in block $b(i)$, T_i is a dummy indicating treatment status, X_i is a vector of household controls, and $\gamma_{b(i)}$ are block-level fixed effects. The coefficient of interest β corresponds to the effect of treatment, and standard errors are clustered at the village level.

This analysis no longer conditions on positive area planted at midline because planting is an endogenous outcome of treatment. Production volume is given by survey response for households with positive area planted and assumed to be zero (though not explicitly asked) for households that previously reported zero area planted.

3 Distribution of Self-Reported Outcomes

In this section we analyze differences in self-reported production by survey mode. This analysis is informative for comparisons made across data sets generated using different methods, for example when making inferences about how outcomes evolve over time from different rounds of a panel survey.

The distribution of responses by survey mode are presented in Figure 1. Each panel plots the value at each decile for the four most common pulse varieties. The solid line represents in-person responses, and the dotted line represents phone responses. Means and standard deviations are also reported for each crop and survey mode.

[Figure 1 about here.]

Data in Figure 1 restrict to study participants that reported positive area planted at midline,

and were therefore asked about production at endline. Nevertheless, some respondents indicate zero harvest production. This is because unfavorable weather conditions in the study year damaged pulse crops, especially pigeon peas. As a result, many households that planted pulses had abandoned cultivation by harvest time.

Results reveal greater self-reported production over the phone than in person. On average, responses range from 8% smaller in-person for fava beans up to 51% smaller for pigeon peas. The difference in means is statistically significant at the 1% level for pigeon peas and at the 5% level for red lentils, the two crops targeted by the extension program. For all crops except fava beans, there is greater variance in responses over the phone as well.

The pattern of greater production reported in phone surveys appears all along the distribution of responses. A larger fraction of respondents claim non-zero production for all crops except for fava beans, and a chi-squared test rejects equality between survey modes at the 1% level for pigeon peas and at the 5% level for green peas. Moreover, self-reported production is higher at the median, 75^{th}percentile, and 90^{th}percentile for all four crops. Differences in pigeon pea responses are significant at the 1% level at the median and $75^{\text{th}}$and at the 5% level at the 90^{th}percentile. Red lentil differences are also significant at the 10% level at the median and 75^{th}percentile, and fava bean differences are significant at the 5% level at the median. Exact values and test statistics are reported in Appendix B. The consistency of these results indicates that the greater mean and variance of phone responses is not just driven by an exaggerated right tail. As a corollary, we would not be able to reconcile survey modes with a simple fix such as more aggressive Winsorization of phone data.

3.1 Selective Attrition and Survey Mode Effects

We first explore differential attrition as a source of difference by survey mode. Table 1 presents household midline characteristics of the 1,533 households that enumerators attempted to contact by phone, which constitute the portion of the sampling frame common to both modes. Column 1 reports means and standard deviations among all households in this population. 715 of these were randomly selected for in-person surveying out of which 702 responded, described in Column 2. Column 3 describes the 1,266 households that responded to the phone survey. Columns 4 and 5 report the in-person and phone sample deviations from the sampling frame, respectively. The top panel reports outcomes asked of all study participants, and the bottom panel reports responses from the extended midline subsample.

[Table 1 about here.]

Attrition was low in person, and endline respondents closely resemble the sampling frame. The only statistically significant deviation is in caste distribution, where there is a slightly lower sampled fraction belonging to a Scheduled Caste or Tribe, almost fully accounted for by Other Backward Castes. All other deviations are quantitatively small and statistically insignificant, consistent with random sampling variation.

By contrast, phone survey respondents appear to be selected along typical dimensions. Households in the phone sample are more educated, with heads four percentage points more likely to have completed primary and secondary school, and appear to be wealthier across a range of measures. Phone respondents are less likely to engage in sharecropping, own more assets, are more likely to live in a permanent housing structure, and are less likely to use government assistance such as workfare (MNREGA) or food aid (PDS). These differences in wealth and education are consistent with selection bias commonly observed in phone surveys.

While the demographic character of phone respondents is associated with greater agricultural output in general, sample selection alone cannot account for measured production gaps between survey modes. To quantify the importance of attrition, we take advantage of the 584 households that responded both in person and by phone, in 429 of which the same individual responded to both surveys. Self-reported production differences within these overlapping subsamples eliminate selection bias and isolate the direct effect of survey mode on the same household or individual responding to the same question over different media.[5]

We first discuss the effect of survey mode on responses given by the same individual. The left column of Figure 2 compares differences between survey modes at each decile among those who responded to both surveys against differences across the full sample of respondents. The solid lines plot the production gap between survey modes at each decile in the full sample, reproducing results from Figure 1, and reflect the net effect of both survey mode and differential selection. The dotted lines represent the production gap in the sample of overlapping respondents, which is only directly affected by survey mode.

[Figure 2 about here.]

For all four main pulse varieties, the production gap at each decile in the overlapping sample closely tracks that of the full sample. The largest deviations occur around the $60^{\text{th}}$to 80^{th}percentiles of green peas, where production reported in person actually exceeds that by phone. Other than this discrepancy, the gap between in-person and phone surveys that appears among the set of respondents who answered both surveys is of similar sign and magnitude to the difference in the full sample throughout the distribution of responses.

Comparing means across the subset of overlapping respondents confirms survey mode effects, rather than selective attrition, generate most of the measured production gap. For three out of four

[5]In Appendix A we show results are not influenced by differences in survey timing.

crops—pigeon pea, red lentil, and green pea—the within-respondent survey mode effect accounts for between 62% and 95% of the average difference across surveys. Only for fava beans is the average production gap within respondent of opposite sign than in the full sample. Figure 2 reveals this reversal is driven largely by a few individuals on the right tail of the production distribution that report substantially more in person.

Within-respondent differences reflect the pure effect of survey mode on the same individual answering the same question. This calculation eliminates heterogeneity caused by both selective attrition across households and by within-household selection of who responds. The latter channel, arising when different members participate in different survey types, can be considered part of the survey mode effect at the household level. Within-household survey mode effects, net of both the direct effect on respondents and household member selection, may be more informative for research design because attrition and the resulting bias can be measured, but researchers cannot know whether the same individual would have responded to different modes of contact.

Household-level survey mode effects explain an even greater portion of the production gap in general. The right column of Figure 2 compares differences in self-reported production by survey mode in the full sample to differences in the sample of overlapping households, represented by the dashed line. The overlapping household sample, consisting of the 429 households in the overlapping respondent sample plus an additional 155 households in which different respondents answered each survey, tracks the full sample more closely across production deciles.

For the two main project crops—pigeon peas and red lentils—the shift from within-respondent to within-household comparisons increases the explanatory power of survey mode effects—from 79% to 89% and 95% to 103%, respectively. Moreover, for fava beans, the portion of the average production gap explained by household-level survey mode effects climbs to 74% with the addition of several households in which the primary farmers reports low production in person and another

member reports higher production over the phone. Interestingly, we document a reversal for green peas as household-level comparisons introduce multiple cases in which the primary farmer reported substantially lower production over the phone than another respondent announced in person.[6]. While effects are not uniformly strong, these results taken together indicate most of the reported production difference between surveys does not come from differential attrition, but rather from the same respondent or household providing different answers based on the manner in which they were contacted.

3.2 Rounding and Respondent Engagement

We next consider differential respondent engagement by survey mode. Phone survey participants may be less engaged for a number of reasons—it is harder for remote enumerators to verify accuracy, it is easier to build rapport face-to-face, or it is more tempting to multitask on the phone, to name a few. Low engagement would add measurement error to survey responses, and may bias responses upward in this context where production volumes are small to begin with.

As a proxy for respondent engagement, we present evidence of rounding in survey responses by plotting the frequency of each value for the right-most digit. Deviations from a smooth distribution, especially around numbers ending in zero and five, would indicate rounding. Gourlay et al. (2019) use crop cuts to show rounding frequently contributes to overestimation of self-reported production data.

Rightmost-digit frequencies are plotted by survey mode and variety in Figure 3. For each crop, we report the fraction of self-reported non-zero production values with each possible right-most digit by survey mode. The figure reveals an excess of responses that end in zero and five. Across all non-zero production data, these two last digits represent 87% percent of responses.

[6]The full breakdown of within-household and within-respondent differences are presented in Appendix B

However, the fraction of responses ending in zero or five is consistent across survey modes. 69% of production values end in zero, 68% over the phone and 70% in person. Similarly, 18% of responses end in five, 17% over the phone and 20% in person. A chi-squared test fails to reject equality in rightmost-digit frequencies at the 5% level, and there is, if anything, slightly more rounding in person. Moreover, the difference is so small that even if rounding caused respondents to halve their self-reported production, it would only lower average production by 1% more in person relative to phone, well below the 8–51% gaps reported in Figure 1. These magnitudes imply that, while participants clearly round their responses, the influence of this behavior on differences by survey mode must be small.

[Figure 3 about here.]

4 Treatment Effect Estimation

Results so far indicate population comparisons between surveys may be undermined by systematic differences caused by survey mode. In this section, we investigate whether cross-sectional relationships within a population remain stable over different modes of survey. This analysis is informative for researchers selecting a method of data collection or comparing results generated using different methods, for example when making inferences about how treatment effects evolve over time within a population.

For this analysis, we report impact evaluation results according to estimation of (1) separately by crop and survey mode. Estimation is straightforward for the in-person sample as it is drawn uniformly at random from the sampling frame. Production quantity is as reported for survey respondents with positive area planted and assumed to be zero for respondents with zero area planted. Regression following (1) produces a treatment effect estimate inclusive of attrition bias

caused by survey non-response.

Comparable estimation in the phone sample is confounded by the fact that enumerators did not attempt to contact households with zero area devoted to pulses at midline. Therefore, the sample consists of a subset of households—those with positive area planted—subject to the attrition pressures induced by phone surveying and a complementary subset—those with no area planted—with known production volume but an unknown phone response rate. These groups are endogenously determined because area planted at midline may be affected by treatment assignment.

To estimate the effect of treatment in the phone sample, we run a weighted least squares regression following (1). Households that responded to the phone survey are assigned a weight of 1, and households with zero area planted are assigned a weight of 0.826 corresponding to the response frequency among surveyed households. Because all non-planting households have an identical production value of zero, this regression recovers the estimated treatment effect inclusive of phone-induced attrition bias under the assumption that phone response rates among non-planting households would have been comparable to response rates among planting households.

Regression coefficients are presented in Figure 4 with 95% confidence intervals subject to a standard error adjustment for sample size. In general, regression standard errors are computed as

$$\sigma_\beta = \frac{\sigma_\epsilon}{\sqrt{N}} \tag{2}$$

a ratio of the residual variance and the sample size, both of which vary by survey mode in our data. In Figure 4, we isolate the residual variance component of (2) by multiplying $\sigma_\beta^{In-Person}$ by $\sqrt{N^{In-Person}/N^{Phone}}$. This correction approximates the regression standard error we would have computed had the in-person survey reached as many respondents as the phone survey while maintaining the same residual variance.[7]

[7] N^{Phone} is calculated as the sum of regression weights. Note this is a simplified approximation because the

[Figure 4 about here.]

Estimated treatment effects are nearly identical in magnitude across survey modes for all four main pulse crops, and a standard t-test fails to reject equality for any crop. This fact remains true even after the $\sqrt{N}$ standard error correction described above, which shrinks the in-person standard errors and thereby raises the probability of rejection. Exact coefficients and standard errors are reported in Appendix B. Notably, the higher attrition rates among pulse producers in the phone survey do not appear to introduce bias. These results indicate that, in contrast to the findings on population moments in the previous section, treatment effect estimation remains stable across survey modes.[8]

While regression coefficients remain stable, Figure 4 shows that standard errors are consistently smaller in the in-person data. This discrepancy highlights a tradeoff in study design: phone surveys, while usually cheaper, generate noisier data. The standard error approximation in (2) provides a straightforward quantification of this tradeoff. To estimate the effect of treatment on pigeon pea production with equal precision, the phone survey would have needed to be 12.2 times larger than the in-person survey; 1.3 times for lentils; 3.1 times for green peas; and 1.6 times larger for fava beans. That is, the cost per response may need to be up to 12.2 times lower over the phone than in person, depending on the outcome of interest, for phone surveying to be a cost-effective method to improve study power.

Estimates in Figure 4 control for household fixed characteristics elicited in person at midline. Dropping these covariates lowers precision, but point estimates remain stable and the relative difference in standard errors persists. This same pattern of consistent point estimates but larger

regression standard errors are clustered at the village level, which is the unit of random assignment to treatment. A more comprehensive correction would need to fully specify differences in the number of clusters, observations per cluster, and intra-cluster correlation by survey mode.

[8]Survey mode may still play a role in the interpretation of outcomes if treatment effects are benchmarked against the control mean or reported in standardized units.

standard errors in phone survey data also appears when restricting to the overlapping subsample of households that participated in both surveys. The implied cost ratio in these specifications leans slightly more in favor of in-person surveying.

5 Conclusion

Overall, this study uncovers meaningful differences in the sample distribution of agricultural output across survey modes. We show a systematic pattern of higher self-reported production over the phone relative to in-person, even within the same respondent, which may bias estimates of local or regional productivity. However, it remains an open question which mode more closely approximates the truth. Validating survey-based production measures would require resource-intensive monitoring of full plot harvests, as self-reporting, remote sensing, and sub-plot crop cuts have all been shown to introduce measurement error (Lobell et al., 2020), and such validation is well beyond the scope of this study.

Our findings more generally highlight a potential concern in interpreting data from national statistical offices. Time series population statistics may be disrupted as survey units update procedures to take advantage of more pervasive information and communication technologies. Changes to aggregation and imputation methods have already proven to generate large discontinuities in historical trends (Jerven, 2013). Survey-mode-induced disruptions may be more difficult to detect when they coincide with technological expansions that cause real deviations from trend, and will be especially obscured where new survey methods were adopted out of necessity during the COVID-19 pandemic. In such cases it will be imperative to design surveys in a way that allows researchers to disentangle true disruptions in underlying outcomes from artifacts of the form of data collection.

Somewhat reassuringly, survey mode effects appear to be less concerning for bias in program evaluation. Gaps in self-reported agricultural production are consistent across experimental study

arms and therefore do not influence the magnitude of estimated program impacts.

Data differences by survey mode are nevertheless important for research design due to precision. We report higher sampling variation in outcome data by phone, though the influence of this difference varies by outcome. In-person surveying at midline further improved precision by allowing us to control for household characteristics. If these covariates were measured more poorly or not at all by phone at midline, the gap in precision between survey modes would have been even greater. Overall, our results caution phone surveying may not save on costs if larger sample sizes are needed to achieve the same level of power.

Implementation experience raises two additional research design considerations not directly quantified in this analysis. First, different survey modes may have different levels of success in reaching specific household members for participation. In our study, in-person enumerators reached the primary farmer slightly more frequently than phone surveyors. Relative success rates may vary across different contexts.

Second, while we focus on the subset of outcomes elicited both in person and by phone, surveys also varied in the scope of their questionnaires. Specifically, enumerators were able to spend over an order of magnitude more time with respondents in person. As a result, in-person surveys generated substantially more data, including production volume for a wider range of crops as well as detailed modules on household income, consumption, and food storage. The ability to reach desired respondents and the breadth of data per respondent add additional dimensions to the tradeoff between cost and precision when selecting a mode of survey for program evaluation.

References

Anderson, Ellen, Rupika Singh, Daniel Stein, and Kate Sturla, "What are the barriers to pulse cultivation in India? Evidence from a randomized controlled trial," Unpublished manuscript 2022.

Biemer, Paul P, "Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing," *Journal of Official Statistics*, 2001, *17* (2), 295.

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On, "Predicting poverty and wealth from mobile phone metadata," *Science*, 2015, *350* (6264), 1073–1076.

Crawfurd, Lee, David K Evans, Susannah Hares, Justin Sandefur et al., "Teaching and testing by phone in a pandemic," 2021.

Deaton, Angus, "Measuring Poverty in a Growing World (or Measuring Growth in a Poor World)," *Review of Economics and Statistics*, 2005, *87* (1), 1–19.

Egger, Dennis, Edward Miguel, Shana S. Warren, Ashish Shenoy, Elliott Collins, Dean Karlan, Doug Parkerson, A. Mushfiq Mobarak, Günther Fink, Christopher Udry, Michael Walker, Johannes Haushofer, Magdalena Larreboure, Susan Athey, Paula Lopez-Pena, Salim Benhachmi, Macartan Humphreys, Layna Lowe, Niccoló F. Meriggi, Andrew Wabwire, C. Austin Davis, Utz Johann Pape, Tilman Graff, Maarten Voors, Carolyn Nekesa, and Corey Vernot, "Falling Living Standards during the COVID-19 Crisis: Quantitative Evidence from Nine Developing Countries," *Science Advances*, 2021, *7* (6).

Ferreira, Aline Dayrell, Cibele Comini César, Deborah Carvalho Malta, Amanda Cristina de Souza Andrade, Cynthia Graciane Carvalho Ramos, Fernando Augusto Proietti, Regina Tomie Ivata Bernal, and Waleska Teixeira Caiaffa, "Validity of data collected by telephone survey: a comparison of VIGITEL 2008 and'Saude em Beaga'survey," *Revista Brasileira de Epidemiologia*, 2011, *14*, 16–30.

Garlick, Robert, Kate Orkin, and Simon Quinn, "Call me maybe: Experimental evidence on frequency and medium effects in microenterprise surveys," *The World Bank Economic Review*, 2020, *34* (2), 418–443.

Gourlay, Sydney, Talip Kilic, and David B Lobell, "A new spin on an old debate: Errors in farmer-reported production and their implications for inverse scale-Productivity relationship in Uganda," *Journal of Development Economics*, 2019, *141*, 102376.

_ , _ , Antonio Martuscelli, Philip Wollburg, and Alberto Zezza, "Viewpoint: High-frequency phone surveys on COVID-19: Good practices, open questions," *Food Policy*, 2021, *105*, 102153.

Greenfield, Thomas K., Lorraine T. Midanik, and John D. Rogers, "Effects of telephone versus face-to-face interview modes on reports of alcohol consumption," *Addiction*, 2000, *95* (2), 277–284.

Greenleaf, Abigail R, Aliou Gadiaga, Georges Guiella, Shani Turke, Noelle Battle, Saifuddin Ahmed, and Caroline Moreau, "Comparability of modern contraceptive use estimates between a face-to-face survey and a cellphone survey among women in Burkina Faso," *PloS one*, 2020, *15* (5), e0231819.

International Telecommunication Union (ITU), "Measuting Digita Development: Facts and Figures 2022," ITU Publications 2022.

Jerven, Morten, *Poor Numbers: How We Are Misled by African Development Statistics and What to Do about It*, Cornell University Press, 2013.

Josephson, A., T. Kilic, and J.D. Michler, "Socioeconomic impacts of COVID-19 in low-income countries," *Nature Human Behavior*, 2021, *5* (1), 557–565.

Kilic, Talip, Heather Moylan, John Ilukor, Clement Mtengula, and Innocent Pangapanga-Phiri, "Root for the tubers: Extended-harvest crop production and productivity measurement in surveys," *Food Policy*, 2021, *102*, 102033.

Lobell, David B, George Azzari, Marshall Burke, Sydney Gourlay, Zhenong Jin, Talip Kilic, and Siobhan Murray, "Eyes in the Sky, Boots on the Ground: Assessing Satellite- and Ground-Based Approaches to Crop Yield Measurement and Analysis," *American Journal of Agricultural Economics*, 2020, *102* (1), 202–219.

Lybbert, Travis, Ashish Shenoy, Tomoé Bourdier, and Caitlin Kieran, "Striving to Revive Pulses in India with Extension, Input Subsidies, and Output Price Supports," Unpublished manuscript 2022.

Maguire, Kelly B, "Does mode matter? A comparison of telephone, mail, and in-person treatments in contingent valuation surveys," *Journal of environmental management*, 2009, *90* (11), 3528–3533.

Mahfoud, Ziyad, Lilian Ghandour, Blanche Ghandour, Ali H Mokdad, and Abla M Sibai, "Cell phone and face-to-face interview responses in population-based surveys: how do they compare?," *Field methods*, 2015, *27* (1), 39–54.

Meager, Rachael, "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments," *American Economic Journal: Applied Economics*, 2019, *11* (1), 57–91.

Nord, Mark and Heather Hopwood, "Does interview mode matter for food security measurement? Telephone versus in-person interviews in the Current Population Survey Food Security Supplement," *Public Health Nutrition*, 2007, *10* (12), 1474–1480.

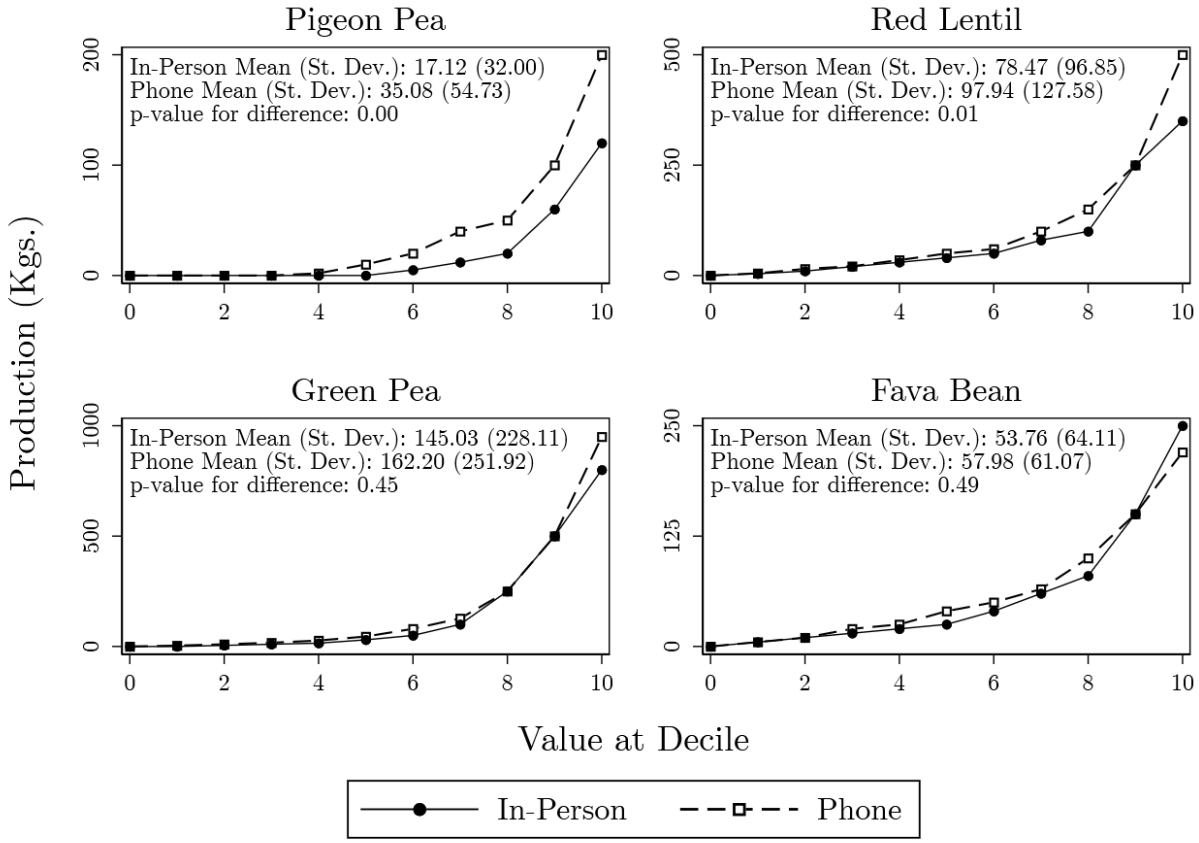
Pritchett, Lant and Justin Sandefur, "Learning from Experiments When Context Matters," *American Economic Review*, 2015, *105* (5), 471–75.

Scherpenzeel, Annette and Philippe Eichenberger, *Mode effects in panel surveys: A comparison of CAPI and CATI*, Bundesamt für Statistik, 2001.

St-Pierre, Martin and Yves Béland, "Mode effects in the Canadian Community Health Survey: A comparison of CAPI and CATI," in "Proceedings of the Annual Meeting of the American Statistical Association, Survey Research Methods Section, August 2004" 2004.

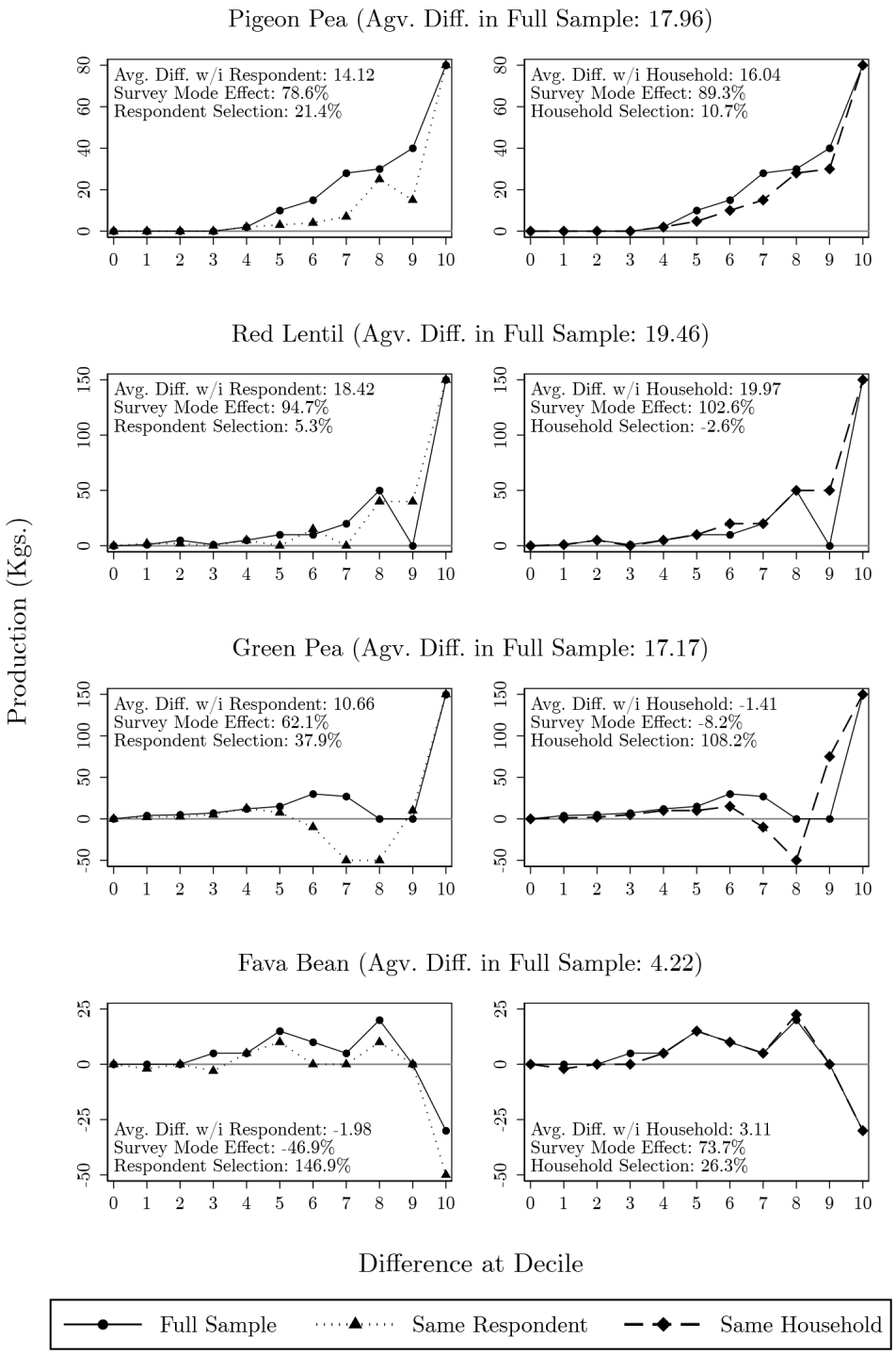
Szolnoki, Gergely and Dieter Hoffmann, "Online, face-to-face and telephone surveys—Comparing different sampling methods in wine consumer research," *Wine Economics and Policy*, 2013, *2* (2), 57–66.

Figure 1: Deciles of Production Quantity by Survey Mode



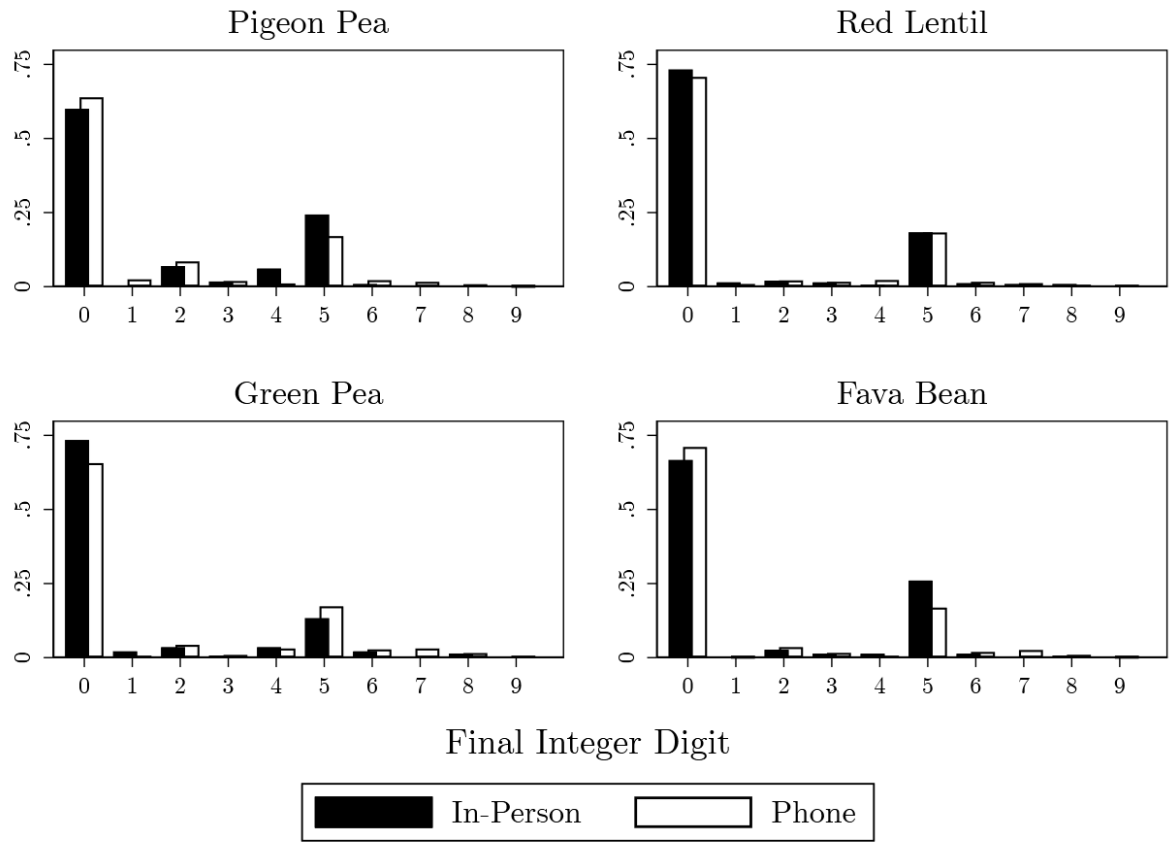
Notes: Self-reported production volume at each decile by crop and by survey mode. Data for each crop includes only those who reported positive area for that crop at midline, and were therefore asked about production of that crop at endline. Top production values are Winsorized to the 95^{th}percentile independently by crop and by mode before computing mean.

Figure 2: Difference at Each Decile in Full and Overlapping Samples



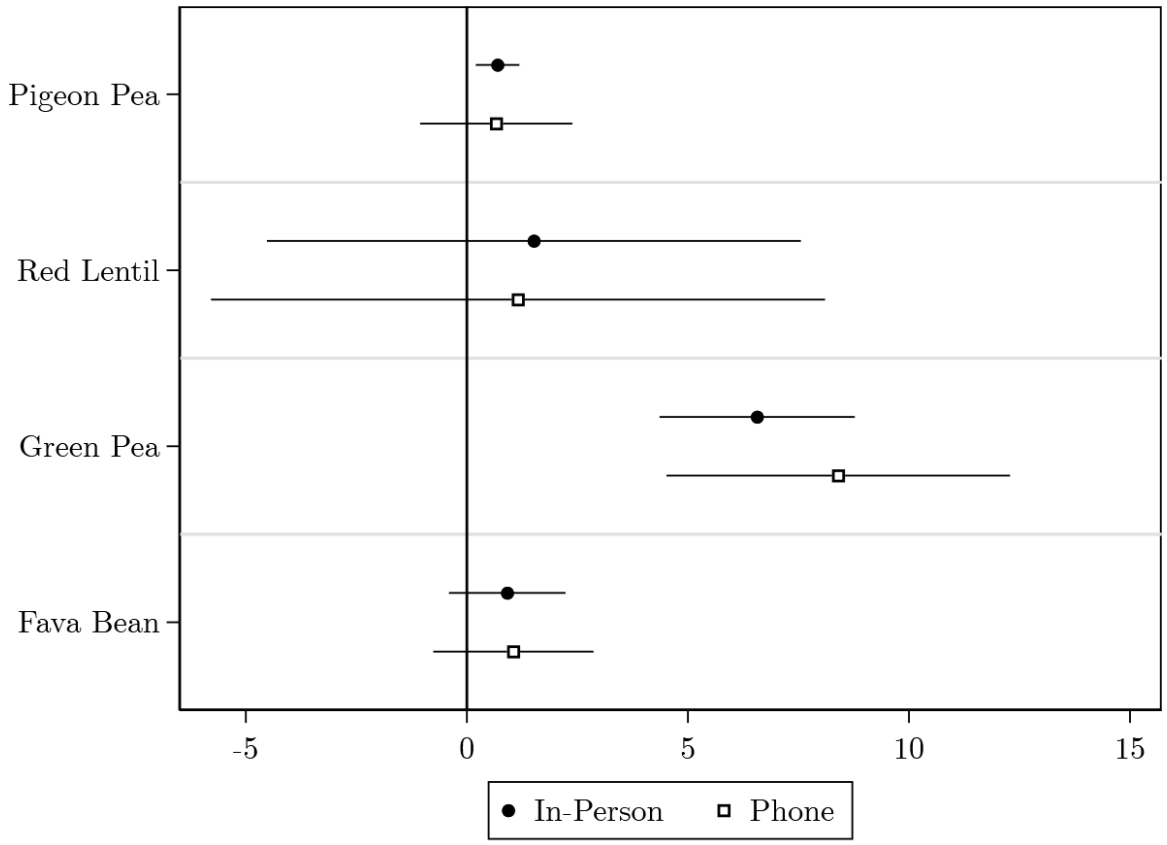
Notes: Difference between self-reported production by phone and in person at each production decile in full and overlapping samples. Data for each crop includes only those who reported positive area for that crop at midline. Left column restricts to overlapping sample with same respondent; right column includes full set of overlapping households. Top production values are Winsorized to the 95^{th}percentile independently by crop and by mode before computing mean difference.

Figure 3: Right-Most Digit Frequencies by Survey Mode



Notes: Fraction of non-zero responses with each value for rightmost digit by crop and by survey mode.

Figure 4: Treatment Effect Estimates by Survey Mode



Notes: Coefficient estimates for treatment effect according to (1) by crop and by survey mode. Error bars represent true 95% confidence intervals for estimation using phone survey data. For estimation using in-person survey data, 95% confidence intervals are shrunk by $\sqrt{N^{In-Person}/N^{Phone}}$ to represent the hypothetical confidence interval had the in-person survey had the same number of respondents as the phone survey. Top production values are Winsorized to the 95^{th}percentile independently by crop and by mode before regression estimation.

Table 1: Household Characteristics by Survey Response Status

	Pulse Growers	Survey Respondents		Difference from (1)	
	Sampling Frame	In-Person	Phone	In-Person	Phone
	(1)	(2)	(3)	(4)	(5)
Variables from full sample:					
HH Head Age	49.125	49.603	49.172	0.479	0.048
	(15.565)	(15.766)	(15.421)	(0.425)	(0.183)
Caste SC/ST	0.167	0.125	0.165	-0.042***	-0.003
	(0.373)	(0.331)	(0.371)	(0.013)	(0.006)
Caste OBC	0.505	0.563	0.506	0.058***	0.001
	(0.500)	(0.496)	(0.500)	(0.016)	(0.006)
Land Farmed (Acres)	2.591	2.470	2.599	-0.121	0.008
	(3.966)	(3.004)	(3.726)	(0.102)	(0.059)
Sharecropping	0.308	0.330	0.295	0.023*	-0.013**
	(0.462)	(0.471)	(0.456)	(0.013)	(0.007)
Observations	1,533	702	1,266		
Variables from detailed subsample:					
Primary School	0.643	0.642	0.681	-0.001	0.038***
	(0.479)	(0.480)	(0.467)	(0.002)	(0.011)
Secondary School	0.482	0.484	0.520	0.002	0.039***
	(0.500)	(0.500)	(0.500)	(0.003)	(0.010)
Asset Index	0.134	0.125	0.258	-0.008	0.124***
	(1.608)	(1.602)	(1.607)	(0.010)	(0.032)
Permanent Housing Structure	0.551	0.548	0.581	-0.003	0.030***
	(0.498)	(0.498)	(0.494)	(0.003)	(0.010)
MNREGA Assistance	0.262	0.258	0.246	-0.004	-0.016*
	(0.440)	(0.438)	(0.431)	(0.003)	(0.008)
PDS Assistance	0.646	0.645	0.626	-0.001	-0.020***
	(0.478)	(0.479)	(0.484)	(0.002)	(0.007)
Observations	715	702	594		

Notes: Household characteristics as reported in the midline survey by endline survey response status. Top panel reports questions asked to all households; bottom panel reports questions asked to extended subsample. Columns 1–3 report sample mean and standard deviation; Columns 4–5 report difference in means from (1) and standard error of difference clustered at the village level. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

Acknowledgements 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 thank Komal Jain, Nandish Kenia, Kate Sturla, and members of IDinsight for study design input and data collection; Tomoé Bourdier and Caitlin Kieran 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 the seminar audience at Northwestern University for 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 Singh and Stein are employed by IDinsight, and author Anderson was employed by IDinsight at the time of data collection. 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. Impact evaluation designs were pre-registered at the AEA RCT registry under AEARCTR-0003872 and the 3ie Registry for International Development Impact Evaluations under RIDIE-STUDY-ID-5a746ac69f12b.

Supplementary Appendix for

"Does Survey Mode Matter?"

For Online Publication Only

A Survey Details

A.1 Survey Timing

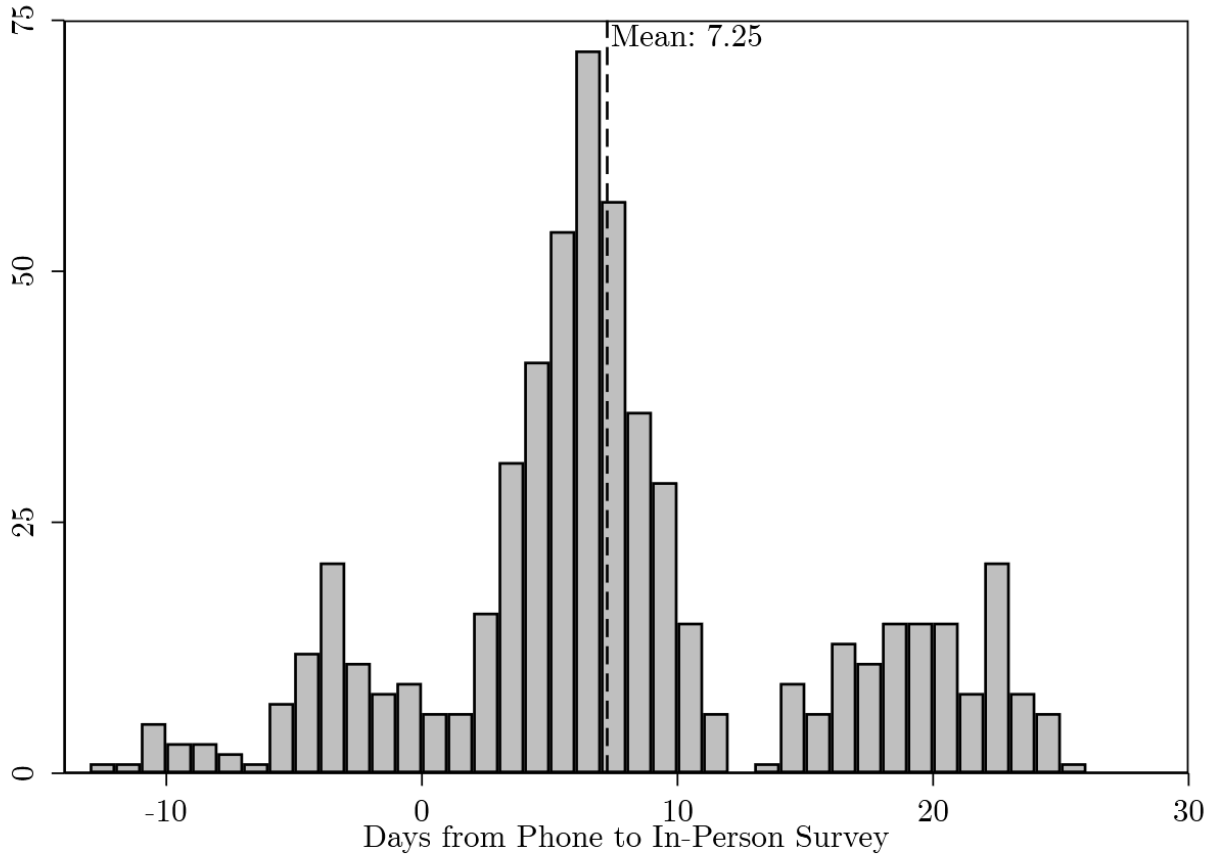
Among the overlapping sample of respondents, the phone survey was conducted an average of one week prior to the in-person survey. Timing differences from the phone survey being conducted 26 days before the in-person survey to 13 days after. The full distribution of the lag between surveys in the overlapping sample is shown in Figure S1.

[Figure S1 about here.]

Survey timing does not seem to affect self-reported production volume. In Figure S2 we plot the production gap between phone and in-person survey responses against the days elapsed between surveys by household for each of the four major crops. In all four cases, there is little relationship between differences in timing and differences in self-reported production, indicating our results are not an artifact of incidental differences in the exact day when respondents were surveyed.

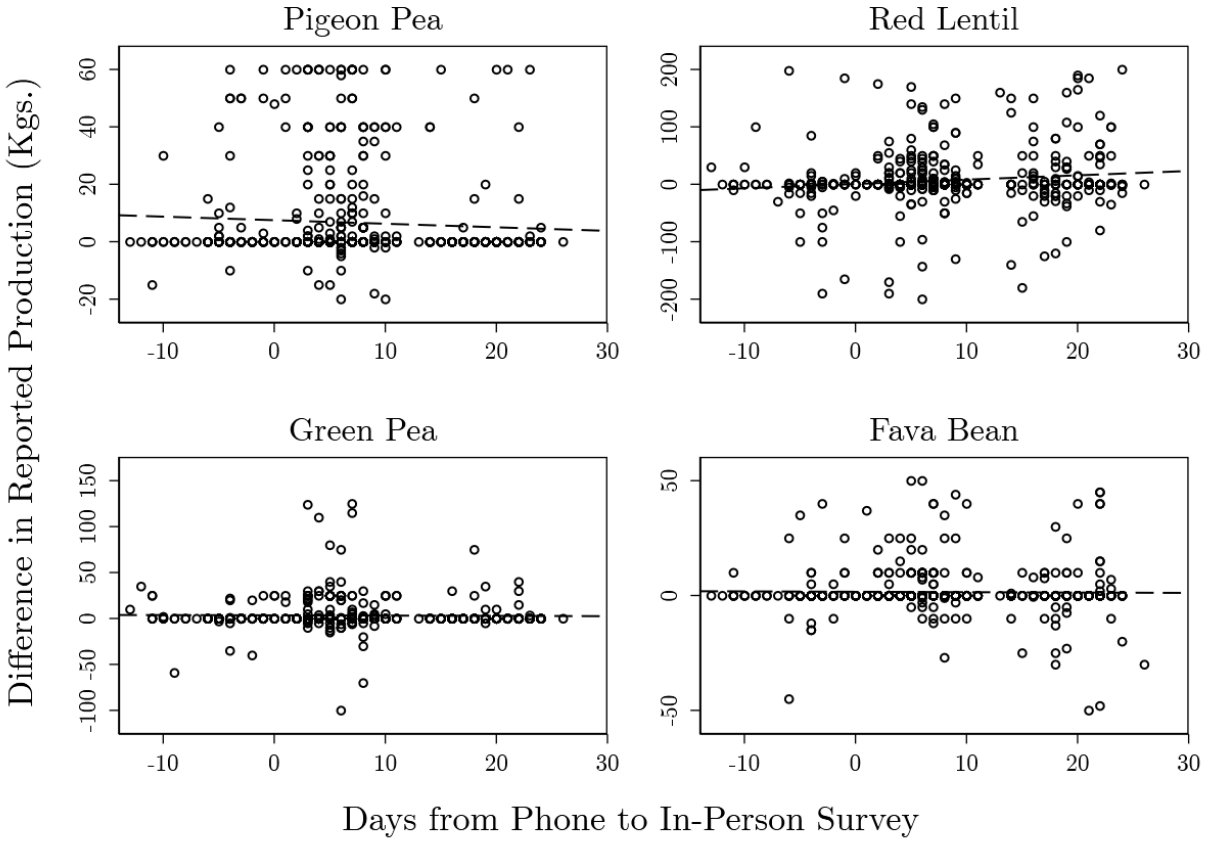
[Figure S2 about here.]

Figure S1: Days Elapsed from Phone Survey to In-Person Survey



Notes: Histogram of days between phone and in-person surveying among households that participated in both surveys. Negative values indicate the in-person survey took place before the phone survey.

Figure S2: Production Gap by Difference in Survey Timing



Notes: Data from households that participated in both surveys only. X-axis represents days elapsed between phone survey and in-person survey. Y-axis represents self-reported production by phone minus self-reported production in person. Dashed line shows best linear fit.

A.2 Survey Questionnaires

Questions on crop production were harmonized between in-person and phone surveys. Excepts containing the exact wording delivered to enumerators for translation into the local language is given below. We show the in-person questionnaire first followed by the phone questionnaire.

Crop name {cropname} will be preloaded from existing data. The number of plots that each crop was cultivated on in Rabi {cropplots} will also be preloaded.

First, I would like to confirm that we have correct information on the crops you cultivated during the last Rabi season. **Ask for each crop that the farmer reported cultivating in the 2017-2018 Rabi season (except orchard crops).**		
B1	You reported cultivating {cropname} during the last Rabi season. Do you confirm that you cultivated {cropname}? Hint: You reported cultivating {cropname} on a total of {cropplots} plots. This includes any {cropname} that you cultivated, for household consumption or commercial purposes, and includes any {cropname} cultivated on the border of your plots.	1. Yes 2. No – Skip to Next Crop

Module C: Production outputs and revenues (Pigeon pea + Rabi crops)

Ask for each crop that the farmer reported cultivating in the 2017-2018 Rabi season (except orchards and any crop that the farmer reported as not cultivating in Module B).

Now we are going to ask some questions about your harvest of {rabi_crop} from all plots.		
C1	How much {rabi_crop} was harvested from all plots in the recent Rabi harvest? Hint: Please enter 0 if {rabi_crop} has not yet been harvested or has only been partially harvested.	
(a)	Quantity: Hint: Please enter 0 if {rabi_crop} has not yet been harvested. Enter -999 if respondent does not know and -998 if respondent refuses to answer.	
(b)	Unit:	1. KGs 2. Quintals 3. Grams 4. Liters 5. Passeri 6. Mann 7. Piece
C2	If the farmer answered 0: Why did you not harvest any {rabi_crop}?	1. Could not afford to cultivate the whole season 2. Insects 3. Rodents/pests 4. Flood 5. Theft 6. Harvest period is later 7. Still harvesting

Section 2: Kharif/Rabi Production			
First, I would like to confirm that we have correct information on the pulses you cultivated during the past Kharif and Rabi season. Then I will ask some questions about your harvest of pulses from all plots. N**ote: Kharif was roughly June-November 2017 and Rabi was November 2017 - April 2018**			
Repeat 2.1-2.12 for each pulse that the farmer reported cultivating.			
2.1	Did you harvest {crop} during this past Kharif or Rabi season?	0. No	
		1. Yes	Skip to 2.4
2.2	Did you grow {crop} in the 2017 Kharif or 2017-2018 Rabi season? When our surveyor came to your household, they asked you about the crops you grew on your agricultural plots. Previously you identified {crop} on the following plots: {landmark1}, {landmark2}, {landmark3}, …	0. No, did not grow	Return to 2.1 for next crop
		1. Yes, grew crop and harvested	Skip to 2.4
		2. Yes, grew crop but did not harvest	
2.3	Why did you not harvest any {crop}?	1. Could not afford to	
		2. Insects	
		3. Rodents/Pests	
		4. Flood	
		5. Theft	
		6. Harvest period is later	
2.8	How much {crop} was harvested from all plots in the most recent harvest?	Quantity: 1. kg	Skip to 2.10
		2. quintal	
		3. gram	
		4. paseri	
		5. maund	
		6. bags	

B Quantitative Results

B.1 Tests of Equality at Quantiles

Comparisons of the fraction of farmers reporting non-zero production and fractions above the median, 75^{th}percentile, and 90^{th}percentile by survey mode for each of the four main crops are presented in Table S1. The third column presents p-values from a ξ^2 test for equality between surveys. To test equality at each percentile, we first calculate the grand value at that percentile across both samples. Then, we test for equality in the fraction of respondents in each survey that report production that exceeds the grand percentile value.

Note that in a few cases, fewer than $100 - N$ percent of respondents report production above the grand N^{th}percentile in both surveys. This is possible because these cases correspond to situations where many responses are bunched exactly at the value at that percentile. We compare the fraction in each sample that report strictly greater production, excluding all those bunched at that percentile value. Conversely, had we compared fractions reporting greater-than-or-equal-to that level of production, such bunching would have generated cases where more than $100 - N$ percent of respondents were counted above the N^{th}percentile in both surveys.

[Table S1 about here.]

B.2 Within-Household and within-Respondent Differences

Figure S3 plots self-reported production over the phone against self-reported production in person by crop for households that responded to both surveys, with shaded dots denoting households in which the same individual answered both surveys.

[Figure S3 about here.]

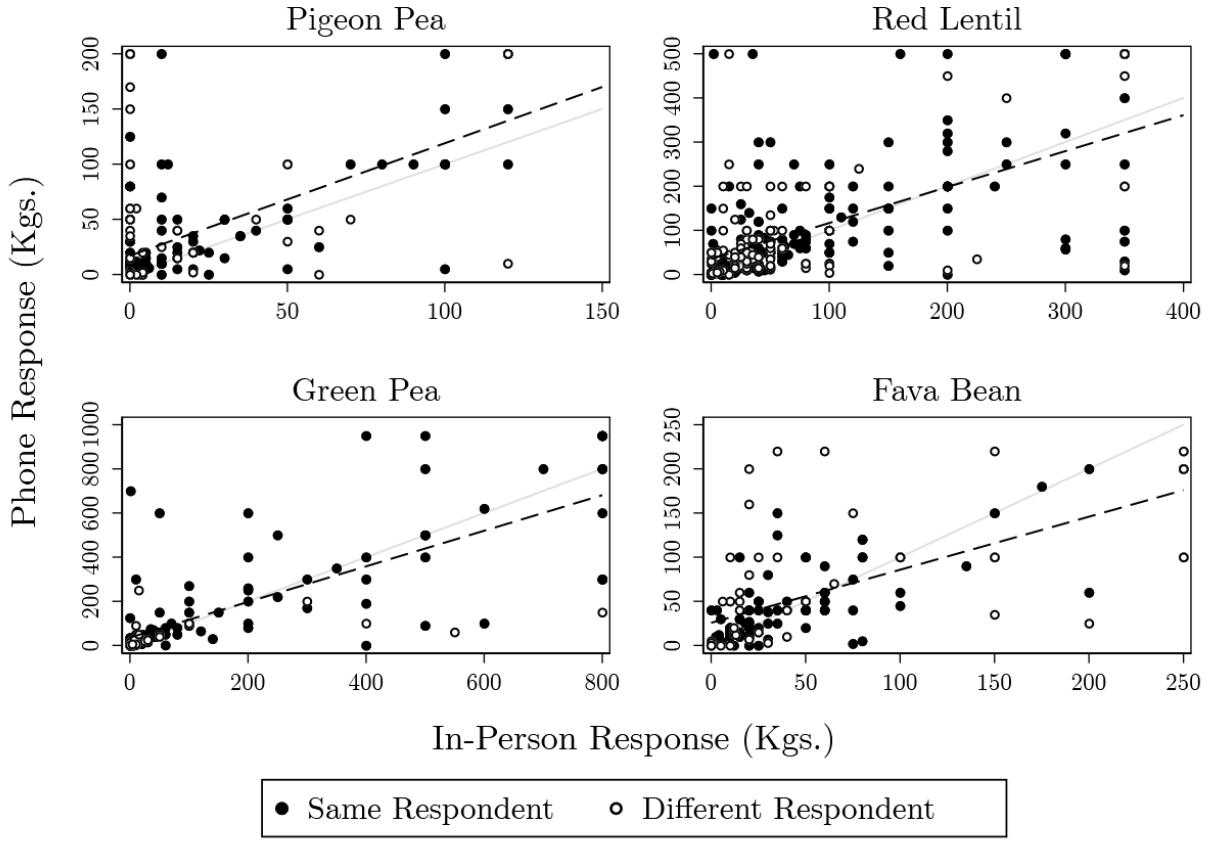
B.3 Program Evaluation Regression Results

Table S2 reports regression results from (1) corresponding to the estimates plotted in Figure 4. Control variables X_i include block fixed effects and respondent age, gender, caste, and experience growing pulses

in prior years, and standard errors are clustered at the village level. Standard errors in Table S2 are not adjusted for sample size differences according to (2).

[Table S2 about here.]

Figure S3: In-Person and Phone Responses by Crop in Overlapping Household Sample



Notes: Production reported by phone plotted against production reported in person by crop among households that answered both surveys. Filled dots represent same respondent; hollow dots represent households with different respondent for each survey. Dashed lines represent best linear fit. Gray lines represent 45-degree line on each graph.

Table S1: Fraction at Various Percentiles by Survey Mode

	Fraction of Respondents		χ^2 Test
	In-Person	Phone	p-value
Pigeon Pea Production:			
Greater than Zero	0.49	0.66	0.00
Above Median	0.38	0.54	0.00
Above 75^{th} Percentile	0.14	0.27	0.00
Above 90^{th} Percentile	0.05	0.09	0.05
Red Lentil Production:			
Greater than Zero	0.93	0.95	0.24
Above Median	0.39	0.45	0.08
Above 75^{th} Percentile	0.20	0.25	0.05
Above 90^{th} Percentile	0.09	0.10	0.61
Green Pea Production:			
Greater than Zero	0.88	0.95	0.01
Above Median	0.44	0.50	0.23
Above 75^{th} Percentile	0.21	0.24	0.49
Above 90^{th} Percentile	0.09	0.10	0.71
Fava Bean Production:			
Greater than Zero	0.94	0.93	0.59
Above Median	0.42	0.52	0.03
Above 75^{th} Percentile	0.19	0.25	0.11
Above 90^{th} Percentile	0.09	0.10	0.75

Notes: The first two columns report the fraction in each survey reporting non-zero production values and fraction in each survey reporting production strictly greater than the 50^{th}, 75^{th}, and 90^{th}grand percentiles across both surveys by crop. The final column reports p-values from a ξ^2 test of equal fractions between surveys.

Table S2: Treatment Effect Estimates: Regression Results

	Pigeon Pea		Red Lentil	
	In-Person	Phone	In-Person	Phone
Treated	0.694	0.665	1.516	1.157
	(0.338)	(0.872)	(4.144)	(3.517)
Control Mean	0.89	5.12	25.16	30.48
R-Squared	0.10	0.11	0.16	0.19
Observations	1055	2066	1055	2079
	Green Pea		Fava Bean	
	In-Person	Phone	In-Person	Phone
Treated	6.566	8.401	0.914	1.052
	(1.516)	(1.967)	(0.905)	(0.918)
Control Mean	2.48	4.87	4.13	5.81
R-Squared	0.13	0.16	0.08	0.10
Observations	1055	2079	1055	2079

Notes: Estimated treatment effect by survey mode following (1). Every specification includes block fixed effects and controls for respondent age, gender, caste, and experience growing pulses in prior years. Standard errors clustered at the village level reported in parentheses. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.