Can survey design reduce anchoring bias in recall data?

Evidence from Malawi*

Susan Godlonton

Williams College, sg5@williams.edu

Manuel A. Hernandez

International Food Policy Research Institute (IFPRI), m.a.hernandez@cgiar.org

Cynthia Paz

International Food Policy Research Institute (IFPRI), c.paz@cgiar.org

Selected Paper prepared for presentation at the 2022 Agricultural & Applied Economics Association Annual Meeting, Anaheim, CA; July 31-August 2

* We thank Michael Murphy, Phoebe Scollard, and Phuong Vo for research assistance, as well as Kelvin Balakasi and Misheck Mphande for dedicated project management. We also thank IPA Malawi and all survey team members who participated in making this project successful.

Abstract

Recall biases in retrospective survey data are widely considered to be pervasive and have important implications for empirical research. In this paper, we leverage the survey design literature and test three strategies to attenuate mental anchoring in retrospective data collection: question order effects, retrieval cues, and aggregate (community) anchoring. We embed a survey design experiment in a longitudinal survey of smallholder farmers in Malawi and focus on anchoring bias in maize production and happiness exploiting differences between recalled and concurrent responses. We find that asking for retrospective data before concurrent data overall reduces recall bias by approximately 34% for maize production, a meaningful improvement with no increase in survey data collection costs. Retrieval cues are less successful in reducing the bias for maize reports, while community anchors can exacerbate the bias. Reversing the order of questions and retrieval cues do not help to ease the bias for happiness reports.

Keywords: Recall bias; mental anchoring; survey design; smallholder farmers; recall data; Malawi

JEL codes: C83; Q12; C93; O13; D19

1. Introduction

Recall biases in retrospective data are well documented (see, e.g., Furnham and Chu Boo (2011) for an extensive review on the topic). In the case of data collection in rural settings, the cognitively demanding processes associated with recalling outputs heightens concerns, particularly among smallholders in developing countries with low levels of education and highly volatile working environments (Arthi et al., 2016; de Weerdt et al., 2016). These concerns are enhanced for crops that constitute a smaller share of household income (Judge and Schechter, 2009) and for outcomes that require eliciting information from many household members or multiple plots (e.g., family farm labor, Arthi et al., 2016). The consequences of these likely mismeasurements have important implications for empirical analyses. For example, recall biases in self-reported data can underestimate income volatility (de Nicola and Gine, 2014), undervalue labor productivity (Gollin et al., 2014), or mischaracterize important relationships such as the inverse size-productivity relationship (Abay et al., 2019). Understanding ways to reduce or bound such biases is thus critical for more accurate and effective research in situations where it is necessary for the researcher to collect more than one round of data at a single point in time.

One distinct source of recall bias commonly documented in rural contexts is anchoring bias, where respondents tend to use their outcome for their most recent period as a cognitive heuristic when recalling the outcome from a previous period (Godlonton et al., 2018). While many innovative data collection strategies are currently being explored to measure outcomes that do not rely on self-reports (e.g., Lobell et al. 2019), most development research is still heavily based on self-reported data and very often on self-reported retrospective data. In this context, further understanding how we can mitigate anchoring bias through survey design is valuable, especially in low-income settings.

In this paper, we leverage the survey design literature and test three alternative strategies to reduce recall error in retrospective data collection: question order effects, retrieval cues (structured prompts), and aggregate (community) anchoring. We base our analysis in a survey design experiment, conducted in 2016, that was embedded in a longitudinal survey (implemented between 2014 and 2018) among smallholder farmers in rural Malawi. Our sample is comprised of cash crop farmers who predominantly grow groundnuts, soy, and tobacco in addition to maize, which is the most important staple crop in Malawi. We focus on reporting biases in one objective and one subjective outcome of interest: maize production and happiness level, respectively. Total maize production is measured in kilograms, and respondents' level of happiness is measured on a ten-point scale. This setup allows us to compare historical concurrent measurements collected during different panel waves with recall measures collected during the survey experiment. Of particular interest to this study are three reports for both maize production and happiness level: concurrent 2015 and 2016 reports measured in 2015 and 2016, respectively, and 2015 retrospective reports measured in 2016.

We purposefully selected an objective and a subjective indicator for the study for two main reasons. First, subjective indicators are increasingly common in the literature but also exhibit lower reliability ratios (e.g., Krueger and Schkade, 2008). Second, the two types of indicators are measured differently and the strategies to mitigate recall bias may operate in a different manner.

We first document the presence of anchoring bias in our setting. Consistent with the broader literature, we find evidence of mental anchoring when collecting retrospective and concurrent information at the same time. Specifically, we find a strong relationship between the 2016 concurrent report and the 2015 retrospective report (collected in 2016), that is also noticeably stronger than the relationship between the 2015 concurrent report and the 2015 retrospective report. This pattern holds for both our objective (maize production) and subjective (happiness) outcomes of interest.

Having established mental anchoring, we then turn to our main contribution that consists in testing the relative efficacy of different survey modules in anchoring bias reduction. First, we assess whether asking about the past before the present can reduce recall bias. We find strong evidence that asking about the retrospective agricultural season before the current season, instead of after, shows substantial improvements in recalling maize production but no impact on reducing recall error in retrospective happiness. We additionally observe that the impact varies across the distribution of the current maize production volume as farmers' most recent production experience continues to play an important anchoring role when recalling past outcomes. For farmers producing lower quantities of maize in the most recent period, the reduction in anchoring bias is most pronounced.

Structured prompts deliberately guide a respondent through a series of related questions in an effort to aid with the memory recall process. Using structured prompts adapted to our outcomes of interest, also meaningfully reduces recall bias (only for maize production) but to a lesser extent than asking about the past before the present. The extent of the reduction is similarly the largest among households producing the least maize in the current period. Finally, we examine community prompts (i.e., an aggregate measure at the community level that can serve as an alternative aggregate anchor). We define community as the farmers' Group Action Committee (GAC), a subunit of the farmer organization to which the farmers belong. This is a well-established administrative structure of the organization and organizes farmers into geographically proximate areas. Using community prompts for maize production is not successful in reducing recall bias. Instead, using such aggregate prompts, whether or not they are paired with retrieval cues, can exacerbate the underlying problem, especially among farmers producing lower quantities of maize in the most recent period.

We further seek to approximate the overall (expected) extent of recall bias reduction across the different survey module designs. This is particularly important given the variation in recall bias across farmer maize production estimates that is mainly driven by their most recent production experience (mental anchor). We accordingly compare median (conditional) deviations of 2015 recalled versus 2015 concurrent maize reports for the different treatment module designs and find that asking about the past before the present, relative to the present before past, translate into a meaningful bias reduction of 34%. This constitutes a relevant and practical finding in that reversing the order of question can reduce recall bias without any increase in survey costs. The structured prompts, in addition to being more costly, are less precise and are expected to reduce recall bias by only 14.5%. Using an aggregate (community) anchor, in contrast, aggravates the bias.

Our paper contributes to the literature on cognitive biases when collecting self-reported recalled data and how can survey design help to mitigate these biases. Recall biases comprise one set of common data quality concerns and previous work has sought to quantify the magnitude of these biases, particularly in agricultural settings, and descriptively characterize recall bias heterogeneity. Beegle et al. (2012) test the importance of recall decay by randomizing households to different lengths of time between harvest and interview months in Kenya, Malawi, and Rwanda. They find limited support for large recall bias over delays in reporting for up to 8 months after harvest; however, statistical power considerations limit the authors in drawing stronger conclusions. Several other papers have found evidence of recall bias in similar settings through different approaches. Arthi et al. (2016) measure family farm labor in Tanzania. Using four different survey design modules they test differences between weekly recall visits (either in person or over the phone) and two different versions of an end-of-season recall module akin to a typical LSMS module. They find compelling evidence that person-plot level family farm labor is over-reported using the end-of season modules; but when aggregated to the household level, this dissipates due to under-reporting of both plots and of family members. The under-reporting of specific household members has also been shown in Ambler et al. (forthcoming), where they document evidence of respondent fatigue for labor reporting as one moves down a household roster. Deininger et al. (2012) experiment in Uganda using agricultural diaries similar in spirit to consumption diaries, which are considered the gold-standard and have shown to yield larger consumption estimates in Tanzania (Beegle et al., 2012). Similar to those consumption findings, Deininger et al. (2012) show that production value using diaries exceeds that of recall-based production modules. Finally, in Central America, Godlonton et al. (2018) examines recall bias among smallholder farmers in four countries in Latin America. In that paper, we find strong evidence of sizeable anchoring bias in self-reported retrospective indicators for both objective measures (income, wages, and working hours) and subjective measures (reports of happiness, health, stress, and wellbeing). Except for the costly innovation of using diaries to help improve agricultural production statistics, the evidence base falls short in offering survey design suggestions on how to lessen the bias.[1] This paper aims to address that gap.

Overall, our results advance this literature offering an actionable recommendation for objective outcomes (such as production) when it is necessary to ask about multiple agricultural seasons retrospectively. Among the menu of survey module options considered, asking about the past before the present is the best approach to limit recall error in this type of setting, where outcomes are typically correlated across time and occur in highly uncertain environments. Farmers tend to use their most recent experience as a cognitive heuristic when recalling past experiences but reversing the order of questions, asking first about the past, can meaningfully reduce this bias.

In this vein, our study also contributes to the active measurement literature in Development Economics that has accelerated in recent years. McKenzie and Rosenzweig (2012) remark that despite the dramatic increase in researcher-led survey efforts globally, little analytical work had been commenced at that time on how to best measure key concepts (outcomes). Since then, researchers have taken on this task and considered measurement challenges for a wide range of outcomes, including use of loans (Karlan and Zinman, 2012), employment (Heath et al., 2021), intimate partner violence (Aguero and Frisancho, forthcoming), among others. de Weerdt et al. (2020) provide a comprehensive overview of the progress made and current best practices that stem from this literature. Our paper intends to contribute to this discussion.

The remainder of the paper is organized as follows. Section 2 describes the experimental design and data. Section 3 details the empirical approach, while Section 4 presents and discusses the estimation results. Section 5 concludes.

[1] Ambler et al. (forthcoming) show that randomized rosters reduce the disproportionate recall error for women and youth, but it does not solve the aggregate problem of respondent fatigue when responding to labor questions as one moves down the household roster.

2. Research Design and Data

2.1 Experimental Design

An extensive interdisciplinary literature has examined survey biases, including biases attributable to the interaction between underlying cognitive processes and the survey instrument (Sudman et al., 1996). We draw on three specific areas from this literature to inform our experimental design: question order effects, retrieval cues, and aggregate anchoring.

Question order effects

The order in which either questions or responses are presented to respondents matters (Schuman and Presser, 1996; Sudman et al., 1996; and Tourangeau et al., 2000). Much of this literature has focused on attitudinal questions (Tourangeau et al., 1991), autobiographical (Tourangeau et al., 2000), or episodic events (Loftus et al., 1992).

The reasoning for why order matters is that a preceding question may have a causal impact on a latter question, which makes it a potential source of recall bias when collecting retrospective data. Some literature has sought to understand the importance of chronology in improving the accuracy of data collected. For example, studies show improvements in survey recall when asking respondents about events in reverse order (starting from the most recent) such as their teachers' names (Whitten and Leonard, 1981) or the topics covered in past psychology exams (Loftus and Farthi, 1985). In both cases, there is no correlation between the correct answers across time. In contrast, in the agricultural domain a correlation does exist where asking about the present first may anchor reports about the past (Godlonton et al., 2018). Given the frequent need to collect retrospective agriculture data, it is not immediately clear what order would minimize recall bias.

Structured prompts

Several researchers have documented that the more time an individual is allocated to answer a question the more accurate the answer is (Tourangeau et al., 2000). Authors have used different strategies to achieve this such as longer questions and/or allocating respondents more time to answer. This is more difficult to control in large (and lengthy) household field surveys, as opposed to lab experiments, especially if payment schemes incentivize a certain number of daily completed surveys combined with other internal survey team pressures. However, one potential way to address this aspect is including additional questions in the survey module in the form of structured prompts that act as memory retrieval cues. The goal is to slow down the memory retrieval process to improve accuracy in data collection.

Aggregate anchoring as a device to improve recall bias

The situation in which individuals use some easily observed prior or recently provided information to guide them in estimating a value under uncertainty (see, e.g., Tversky and Kahneman, 1974) has been widely documented in both the economics and survey methodology literature.[2] In the case of retrospective agricultural data collection, farmers are generally influenced or mentally anchored by their current agricultural season experience (as in our sample). We alternatively seek to use an aggregate anchor (average community production in the past season), independently as well as combined with

[2] Some examples include Hurd et al. (1998), Frykblom and Shogren (2000), Ariely et al. (2003), Beggs and Grady (2009), Campbell and Sharpe (2009), and Hitczenko (2013).

memory retrieval cues, to elicit more accurate retrospective data; by explicitly using an aggregate (community) anchor about production in the past season we intend to disrupt the mental anchor and replace it with a more proximate anchor.

Experimental design

Using these theoretical insights, we design a survey experiment to test the effect of different survey module features on recall bias reduction. We focus on both an objective outcome (maize production, measured in kilograms) and a subjective outcome (happiness, measured on a ten-point scale). As noted earlier, the interest in both an objective and subjective indicator is the increasing use of subjective data in the economics literature (besides quantitative data) and the fact that subjective measures typically exhibit a lower reliability than objective measures. The approaches to attenuate recall bias may also work differently between objective and subjective indicators.

The survey experiment was conducted during the second follow-up survey of a randomized controlled trial agricultural intervention, discussed in detail in Ambler et al. (2018a). The goal of the agricultural intervention was to boost productivity among smallholder farmers using various framed capital transfers and modalities of extension services. The broader project was conducted in the central region of Malawi, specifically in Dowa and Ntchisi districts, between 2014 and 2018. The sample includes all households associated with the 120 newest farmer groups registered with the National Smallholder Farmers Association of Malawi (NASFAM). Ambler et al. (2018b) show that the households included in the study are comparable to other agricultural households in Dowa and Ntchisi districts. The main crops grown in the area include maize (almost universally grown), groundnuts, soy, and tobacco. While tobacco has been the primary cash crop in this region for decades and many households within our sample continue to grow tobacco, the specific farming groups targeted for project participation were groundnut and soy farmer groups. We focus on maize production in this study to preserve statistical power for the survey design experiment as maize is the main staple crop in the region (and country) grown by most sampled farmers.

The treatment conditions vary along three dimensions: the sequencing of retrospective and concurrent data collection; whether or not we use structured prompts during the retrospective data collection process; and whether or not we include a community anchor (using aggregate data) to frame respondents regarding the previous agricultural season prior to eliciting retrospective production data. Using different combinations of these three elements, we develop a survey experiment with six distinct treatment conditions (survey module designs) that were randomly assigned to respondents. The design is purposefully a non-nested design. A nested design would suffer from anchoring biases of one treatment condition affecting a latter treatment condition.

To test whether asking first about the past can reduce recall bias we implement the following two treatment conditions:

Treatment Condition 1. Respondents are asked about their maize production and level of happiness in the current year (2016), and then asked about both in the prior year (2015).

Treatment Condition 2. This treatment reverses the order. Respondents are first asked about maize production and happiness in the previous year (2015), and then asked about both in the current year (2016).

Similarly, to test whether structured prompts have potential to reduce recall bias we implement the following third treatment arm:

Treatment Condition 3. Respondents are asked about their maize production and happiness in the current year. Then, we use individual framing questions before eliciting their maize production and happiness in the prior year. The individual prompts first ask the respondent whether production or happiness this year (2016) is about the same, greater, or lower than in the previous year (2015); then, they ask whether it was much higher/lower, moderately higher/lower, or just a bit higher/lower. Only after this exercise, respondents are asked about their production or happiness in the previous year. The idea is to force more deliberation that could aid respondents better recall their last year's outcome.

Finally, to test whether community anchors can reduce recall bias we add the following three treatment conditions, which is only applicable to maize reports:

Treatment Condition 4. Respondents are only asked about the quantity of maize produced in the prior year (in 2015). This treatment is included in the design solely for comparison purposes as it only asks about the past and not the present and, in most circumstances, this would not be aligned with researchers' intentions.

Treatment Condition 5. Respondents are first informed what the average community production was in their local area in 2015. Then, respondents are asked about their own past maize production. In designing the appropriate community prompt, we opted for the farmers' Group Action Committee (GAC) as the designated community group. These committees are subunits of NASFAM that are spatially determined and are part of the administrative structure of the organization. Individuals within the same GAC will face similar weather conditions and all farmers within the GAC are cash crop farmers. Thus, while farmers within GACs will be heterogeneous in their outputs due to land size, labor endowments, and a host of other factors, the GAC prompt should provide a reasonable reference point to the farmer. Furthermore, administrative data is often summarized by GAC, thus making historical records easily accessible, and thus a feasible option to scale if useful. In sum, the community anchor we use is the average production within the respondent's GAC based on (concurrent) self-reported data collected in 2015. The idea is to deliberately frame the respondent based on a comparable aggregate measure that could help respondents better recall their past outcome.

Treatment Condition 6. Similar to Treatment Condition 5, respondents are first informed about the community level production in the previous year (2015). Then, we ask respondents whether their production in the previous year was higher, lower, or about the same as others in their community as well as whether it was much higher/lower, moderately higher/lower or just a bit higher/lower than their community in the previous year. After this, respondents are asked to recall their production (or happiness) in the previous year.

Hence, Treatment 4 provides a comparison against which we measure the efficacy of Treatments 5 and 6, where treatment Group 5 relies on community level anchors and Group 6 combines community level anchors with structured prompts.

The specific formulation of the questions for each treatment condition is detailed in Appendix A and Table 1 summarizes the sample size and key characteristics of the treatment arms. These treatment survey module designs were purposefully selected to test low-cost alternatives to improve the quality of retrospective data collection efforts. While changing the order of questions adds no cost to data collection in terms of survey time, the added cost of using structured prompts and aggregate anchoring priming is on average 24 and 32 seconds per retrospective question, respectively. Overall, using these treatment conditions allows us to quantify whether and to what extent: asking about the past before the present (Treatment 1 versus Treatment 2), structured prompts (Treatment 1 versus Treatment 3), and using community anchors without and with prompts (Treatment 4 versus Treatment 5 and Treatment 6) are effective at reducing recall bias.

2.2 Data

Figure 1 outlines the timeline of activities most relevant to this paper. The survey experiment is implemented at the end of the second Midline Survey in 2016, but we use data collected in earlier rounds for the broader project (2014 Baseline Survey and 2015 first Midline Survey). Annual surveys were implemented that were reasonably comprehensive in content, modelled largely based on the Integrated Household Survey (IHS) carried out by the National Statistical Office of Malawi to monitor household socioeconomic conditions.[3] From the baseline survey conducted in 2014, we use baseline characteristics to demonstrate balance of the treatment conditions as well as controls in our specifications. From the first midline survey conducted in 2015, we primarily use the concurrently measured outcomes of our key dependent variables: maize production and level of happiness. The focus on the survey experiment in the second midline survey consists of examining bias reductions of different survey designs to accurately collect 2015 data in 2016.

For purposes of this study, we restrict our sample to households who completed both the first and second midline survey, including the module with the experimental survey design. Our working sample comprises 967 households, which include 941 maize producers and 964 households that report their happiness level.[4] The primary respondent is the NASFAM member in the household, who need not necessarily be the household head.

Table 2 presents basic descriptive characteristics of our working sample regarding individual and household characteristics (Panel A), maize production (Panel B), happiness level (Panel C), and interview characteristics (Panel D). We observe that household heads are primarily male (84.9%) and older (average age: 45), while respondents are commonly female (64%) and slightly younger (average age: 42). Households have on average 5.6 household members and own roughly 4.7 acres of land. Using concurrent production reports among maize producers in the sample, the average maize production was 1,055 kilograms in 2015, while in 2016 it was markedly higher averaging over 1,400 kilograms. This sharp increase in production is driven by several factors including more favorable weather conditions for the 2016 harvest as well as overall positive production treatment effects attributable to the broader ongoing agricultural transfer program the broader study was evaluating.[5] The concurrent happiness levels also

[3] The structure of the IHS generally follows a typical Living Standards Measurement Study and includes modules eliciting respondents' demographics, income and labor activities, assets and expenditures, detailed agricultural production information, among others.
[4] Three households that report producing maize did not report their happiness level.
[5] We similarly observe an increase in maize production across all treatment groups except Group 4.

show an increase, although marginal, between 2015 and 2016. The treatment survey modules took on average about 11 minutes.[6]

Using baseline characteristics, we test whether the random assignment to different survey modules is balanced. Appendix Table B.1 presents the corresponding results. First, we run a multinomial logistic regression and test for joint orthogonality across the six treatment conditions (survey module designs) controlling for an extensive set of regressors including individual, household, and interview characteristics. We separately perform this test for the estimating sample used in the maize production recall analysis (941 observations) and the happiness recall analysis (964 observations). The p-values associated with the joint orthogonality test (reported in Panel A of the table) are 0.237 and 0.152 in the maize and happiness subsamples, respectively, which support that our treatment groups are balanced.

Second, we conduct a joint orthogonality test across treatment conditions for each individual (household) and interview characteristic. As observed in Panel B of the table, we find that most covariates are generally balanced across treatment groups. There are only three characteristics (out of 17) that in at least one of our samples, there appears to be some imbalance at conventional statistical levels: whether the household head completed primary school; the duration of the experimental survey module design; and household size (in the maize sample) and the district in which the household resides (in the happiness sample). As described earlier, we expect differences in duration across modules due to the different structure of the questions. The imbalance in the other indicators is mainly driven by Treatment Groups 4 and 5 where a lower share of respondents fall into the completed primary schooling category and are also less likely to reside in Dowa district compared to the other treatment groups. Treatment Group 6 exhibits slightly larger household sizes relative to all other groups. These imbalances, however, are not due to non-compliance as the randomization across treatment groups was performed ex-ante and only the relevant experimental version of the survey module appeared for each corresponding respondent. Our main results are also not sensitive to the exclusion of controls.

3. Empirical approach

Before assessing which of the survey designs is most effective at reducing recall bias in retrospective data collection, we first establish the extent of mental anchoring among farmers in self-reported retrospective measures within this study. In line with Campbell and Sharpe (2009) and Godlonton et al. (2018), we test for the presence of anchoring bias when recalling maize production (objective measure) and the level of happiness (subjective measure) estimating the following linear regression model,

$$y_i^{t-1,t} = \alpha_1 + \alpha_2 y_i^{t-1,t-1} + \alpha_3 y_i^{t,t} + \alpha_4 X_i + \alpha_5 Z_i + \varepsilon_i \qquad (1a)$$

where $y_i^{t-1,t}$ is the recalled measure for the variable of interest (i.e., maize production in kilograms or level of happiness on a scale 1-10) that farmer i reports at current period t in reference to prior period

[6] The treatment modules also included recall questions for groundnuts and soy, which are not part of this study as discussed above.

$t-1$ (e.g., maize production for 2015 reported in 2016); $y_i^{t-1,t-1}$ is the concurrent report of the variable for period $t-1$ reported at $t-1$ (e.g., maize production for 2015 reported in 2015 or 2015 concurrent report); $y_i^{t,t}$ is the concurrent report of the variable for period t reported at t (e.g., maize production in 2016 reported in 2016 or 2016 concurrent report); X_i is a vector of controls that includes individual and household characteristics measured at baseline, as well as district location; Z_i is a vector of interview characteristics; and ε_i is a white noise error.

The parameter of interest in model (1a) is α_3, which captures the partial correlation between the 2016 concurrent report (i.e., the potential anchor value) and the 2015 recalled measure. If farmers use their reported measure for the most recent period as an anchor when recalling the measure for the previous period, then $\alpha_3 > 0$; otherwise, $\alpha_3 = 0$. Parameter α_2 captures, in turn, the partial correlation between the 2015 concurrent report and the 2015 recalled value. The 2015 concurrent report serves as a reference value for the 2015 "true" outcome and helps to account for the correlation in the reported outcomes over time that would otherwise be embedded in parameter δ.[7] As both the 2015 and 2016 concurrent reports may still be subject to measurement error, an implicit assumption in the regression setting is that any measurement error in the reported measures is orthogonal to the anchoring behavior (bias).

We similarly analyze the magnitude and direction of the potential anchoring bias by estimating the following alternative model in deviations,

$$\triangle\, y_i^{t-1} = \beta_1 + \beta_2 \triangle\, y_i^{t} + \beta_3 X_i + \beta_4 Z_i + \varepsilon_i \qquad (1b)$$

where $\triangle\, y_i^{t-1}$ is the difference between the 2015 recalled measure and the 2015 concurrent report for the variable of interest (e.g., the difference between maize recalled production for 2015 reported in 2016 and maize production for 2015 reported in 2015); and $\triangle\, y_i^{t}$ is the difference between the 2016 and 2015 concurrent reports (e.g., the difference between 2016 and 2015 maize production reported in 2016 and 2015, respectively). The parameter of interest in this case, β_2, captures the partial correlation between differences in the 2016 and 2015 concurrent reports and deviations in the 2015 recalled and concurrent reports. If anchoring bias plays an important role when recalling information, we would expect $\beta_2 > 0$. That is, if a farmer reports a larger production in 2016 relative to 2015, we expect a positive deviation in the 2015 recalled value relative to the 2015 concurrent report; and the opposite if the farmer reports a smaller production in 2016 relative to 2015. If anchoring bias does not play any role when recalling information, $\beta_2 = 0$.[8]

After documenting evidence of anchoring recall bias, we then turn to our key study contribution, and assess how farmers' recalling and anchoring behavior varies with the different survey module designs (treatment groups) detailed in Section 2. In particular, we separately compare the behavior of farmers when the order of questions regarding concurrent and retrospective measures are reversed (i.e., Treatment 2 versus Treatment 1); when adding structured prompts to collect retrospective data (i.e., Treatment 3 versus Treatment 1); and when using aggregate community (group) data to frame the past

[7] The partial correlation between the 2015 and 2016 concurrent reports is 0.2 for maize and 0.16 for happiness.
[8] Note that while models (1a) and (1b) are related, they do not produce similar results unless we further impose that $\alpha_2 = 1 - \alpha_3$.

season prior to directly collecting retrospective measures with and without structured prompts (i.e., Treatments 5 and 6 versus Treatment 4).

We accordingly augment equation (1a) to,

$$y_i^{t-1,t} = \alpha_1 + \alpha_2 y_i^{t-1,t-1} + \alpha_3 y_i^{t,t} + \rho_k \, T_{ik} + \delta_k \, T_{ik} * y_i^{t,t} + \alpha_4 X_i + \alpha_5 Z_i + \varepsilon_i \quad \text{(2a)}$$

where T_{ik} is a binary indicator that takes the value of one if farmer i is assigned to treatment module k, and zero otherwise. For example, when comparing Treatment 2 versus Treatment 1 group, binary indicator T_{i2} takes the value of one if farmer i responded to Treatment 2 module (i.e., first recalled their 2015 outcome and then reported their 2016 outcome), and zero if farmer responded to (base) Treatment 1 module (i.e., first reported their 2016 outcome and then recalled their 2015 outcome). We do the same when comparing Treatment 3 and Treatment 1 groups (using binary indicator T_{i3}) and when comparing Treatments 5 and 6 with Treatment 4 group (using binary indicators T_{i5} and T_{i6}).

The parameters of interest in equation (2a) are ρ_k and δ_k. Parameter ρ_k is the shifter that captures the average difference in recalled values between treatment group k and the base treatment group, while parameter δ_k captures the difference in the degree of anchoring between the two groups. The overall difference in recalling behaviors between groups ultimately depends on the magnitudes and signs of parameters ρ_k and δ_k as well as on the reported anchor value (i.e., 2016 concurrent report)

We similarly analyze the difference in recalling and anchoring behavior across treatments by augmenting model (1b) in deviations to,

$$\triangle \, y_i^{t-1} = \beta_1 + \beta_2 \triangle \, y_i^t + \gamma_k T_{ik} + \theta_k \, T_{ik} * \triangle \, y_i^t + \beta_3 X_i + \beta_4 Z_i + \varepsilon_i \qquad \text{(2b)}$$

where the parameters of interest are γ_k and θ_k. Parameter γ_k measures the average difference of the deviation between the 2015 recalled and concurrent reports for treatment group k versus the base treatment group, while parameter θ_k measures the difference between the two groups in the degree of association between deviations in 2016 and 2015 concurrent reports and deviations in 2015 recalled and concurrent reports. The overall difference in recalling behaviors between groups in this case depends on the magnitudes and signs of parameters γ_k and θ_k as well as on the difference between the 2016 and 2015 concurrent reports.

4. Results

We first briefly document the presence and magnitude of mental anchoring in self-reports of maize production and happiness under recall. While an extensive literature has previously shown mental anchoring to be a problem, establishing it exists in the current context is necessary before testing strategies to address it. We then examine whether the different survey module designs described above contribute to reduce anchoring behavior and compare their extent of bias reduction (if any). Finally, we assess whether the results are driven by a specific group of farmers.

4.1 Anchoring behavior

Table 3 reports the results of the regression model where we first quantify the existence of mental anchoring among farmers in self-reported retrospective outcomes. The first four columns correspond to the model for maize production (in kilograms) and the last four columns to the model for level of happiness (on a scale 1-10). Odd-numbered columns do not include any controls, while even-numbered columns add a set of individual (respondent) and interview characteristics. For each outcome, the first two specifications correspond to specification (1a), the model in levels; and the second two specifications correspond to regression equation (1b), the model in deviations. The individual controls include age, gender, and level of education of the respondent, age, and gender of household head, if household is polygamous, household size, plot size, and district location, all collected during the 2014 baseline survey. Interview controls include the gender of enumerators in all survey rounds and duration of the 2015 and 2016 surveys as well as the duration of the recall module implemented in the 2016 survey.[9] The standard errors reported in parentheses are heteroskedastic robust.

As shown in the table, we find evidence of mental anchoring in self-reported recalled measures. For the full specification in levels reported in columns (2) and (6), the coefficient of the (anchoring) 2016 concurrent report is 0.169 for maize production and 0.197 for happiness. Hence, all else equal, an additional kilogram of maize in the 2016 concurrent report is associated with a 0.169-kilogram increase in the production recalled for 2015, while a one-point increase in the 2016 concurrent happiness level is associated with a 0.197-point increase in the recalled happiness level for 2015. The 2015 concurrent report, in turn, is only partially correlated with the 2015 recalled value for maize in the order of 0.12, which is somewhat smaller than the correlation between the anchor and recall measure, and for happiness there is no correlation between these two reports.[10] For the specification in deviations in columns (4) and (8), the apparent mental anchoring is more evident. In the case of maize, a one-kilogram increase in the difference between the 2016 and 2015 concurrent reports is correlated with a 0.42-kilogram larger difference between the 2015 recalled and concurrent reports; for happiness, a one-point increase in the difference between the 2016 and 2015 concurrent reports is correlated with a 0.616-point larger difference between the 2015 recalled and concurrent reports.

The results confirm sizeable mental anchoring in self-reported retrospective measures for both objective (production) and subjective (happiness) outcomes over relatively short (one year) recall periods. The findings are in line with the anchoring bias documented by Godlonton et al. (2018) among smallholder farmers in Central America, although the magnitude of the bias is smaller in this case but in the range of

---

[9] We also include separate indicator variables that take the value of one for missing values on individual or household characteristics, plot size, and survey duration (in these cases we assign the median district value), and zero otherwise. For surveys with implausible minimum or maximum duration values (e.g., surveys partially recorded or paused and re-commenced), we assign the corresponding minimum or maximum duration taken by a regular survey (at the district level) and include an indicator variable equal to one for those cases, and zero otherwise.

[10] If we exclude the 2015 concurrent report from the regressed model, the coefficient for the (anchoring) 2016 concurrent report varies slightly for both maize and happiness, which points to the lack of influence of the 2015 concurrent report on the strong association between the 2016 concurrent report and 2015 recalled value.

studies applied in other contexts (e.g., Beggs and Graddy, 2009; Campbell and Sharpe, 2009).[11] The results of the model in deviations further indicate that positive and negative changes in the concurrent reports between periods appear to influence the anchoring bias in the same direction, which are tendencies that have been widely documented in the psychology literature (e.g., Baumeister et al., 2001; Walker et al., 2003) as well as in Godlonton et al. (2018). Overall, the presence of mental anchoring motivates considering alternative survey designs that can help reduce this bias, which are discussed next.

4.2 Alternative survey designs to attenuate anchoring bias

We now turn to assess how different survey module designs can help to reduce mental anchoring among farmers when collecting self-reported retrospective data. These include switching the sequencing of questions regarding concurrent and retrospective data, using structured prompts during retrospective data collection, and alternatively using aggregate (group) data to frame the previous agricultural season prior to collecting retrospective data without and with structured prompts. We rely on the regression framework described in equations (2a) and (2b) where we compare self-reported retrospective reports (in levels and deviations) obtained across the different treatment groups.

4.2.1 Past before present

We first examine the case where we switch the order of questions to evaluate if the most recent 2016 outcome continues to be a strong anchor for the 2015 recalled value if the question order is inverted. Table 4 reports the results comparing the behavior of respondents who were asked to first recall their 2015 outcome and then to report their 2016 outcome (Treatment 2) versus respondents who were asked to first report their 2016 outcome and then recall their 2015 outcome (base Treatment 1). All regressions hereafter control for the full set of respondent and interview characteristics as well as for the corresponding 2015 and 2016 concurrent reports (or their difference for the model in deviations).[12]

Several interesting patterns emerge from the table. While for maize we observe differences in the recalled reports between reversing and not reversing the order of questions, for happiness the ordering of questions does not seem to matter. Asking first about the past (Treatment 2) results in a lower 2015 maize recalled value and a lower deviation of the recalled value from the value reported in the previous period, compared to asking first about the present. Yet, the volume of production in the most recent period is still relevant and appears to attenuate this effect as noted in columns (2) and (4). A higher 2016 concurrent report or a larger positive difference in the 2016 versus 2015 concurrent report lessens (and eventually reverses) the negative effect of switching the order of questions, and vice versa.

To better appreciate these counteracting effects, Figure 2 Panel A reports the estimated effects on 2015 maize recalled values from reversing the order of questions for different percentile values of 2016

[11] Godlonton et al. (2018) focus on income, wages, and working hours that might require additional mental efforts and calculations among farmers (than recalling production volumes as in this study), which could explain the larger anchoring coefficients for their analysis in Central America. Contrary to our findings, they find larger reference dependence for objective outcomes than for subjective outcomes such as happiness.

[12] The results without controls are though qualitatively similar.

concurrent reports (10, 25, 50, 75, and 90 percentiles).[13] We observe that the negative effect on maize recalled values when asking first about the amount produced in the previous period (relative to asking first about the most recent period) decreases and ultimately reverses as the amount produced in the most recent period increases. For example, for farmers reporting to have produced 250 kilograms of maize in 2016 (10 percentile), their 2015 recalled production value is roughly 349 kilograms lower when asking first about the past (compared to asking first about the present), while for farmers reporting to have produced 2,880 kilograms of maize in 2016 (90 percentile), the 2015 recalled value is 345 kilograms higher. For the sample median of 1,000 kilograms, the effect of asking first about the amount produced in 2015 (instead of 2016) is negative and in the order of -151 kilograms.[14] A similar pattern is observed for the model in deviations reported in Appendix Figure B.1 Panel A.

Hence, asking first about the outcome for the previous period and then to report the outcome for the most recent period can help to reduce mental anchoring among farmers when collecting retrospective production data. The attenuation, however, is conditioned by the amount produced in the most recent period. That is, regardless of the sequencing of questions, the production volume (performance) experienced by farmers in the most recent period continues to play an important role when recalling their past production volume. In our sample, for 2016 concurrent reports roughly below the median value the estimated effect of switching the order of questions is negative at conventional levels, while for values above the 90 percentile is positive.

It is worth noting that we additionally checked that reversing the order of questions, i.e., asking first about the past and then about the present, does not result in anchoring the 2016 concurrent value (reported later) with the 2015 recalled value (reported first). We regressed the 2016 concurrent value on the 2015 recalled value, the dummy variable distinguishing Treatment 2 from Treatment 1, the interaction of the 2015 recalled value with the treatment dummy, and all controls depicted in Equation (2a), including the 2015 concurrent value. We do not find evidence of differential effects of 2015 retrospective responses on 2016 concurrent responses between the two treatment groups.[15]

4.2.2 Using structured prompts

A second alternative survey module design preserves the order of questions, asking first about the most recent outcome but adds structured prompts prior to asking about the past outcome. The idea is to assess whether introducing a structured prompt, which deliberately slows down the interviewees' response and is intended to help them mentally process their answer, can aid farmers in recalling their previous outcome more precisely.

Table 5 presents the results comparing the behavior of these group (Treatment 3) relative to the group who were asked to report the most recent outcome and then to directly recall their outcome from the previous period (base Treatment 1). We find a similar pattern as when reversing the order of questions.

[13] The vertical lines in the figure are the 95% confidence intervals of the reported point estimates derived using the delta method.

[14] For the sample mean of 1,406 kilograms, the effect is -44 kilograms. We prefer to focus on the midpoint or median effect (and percentile effects) for inference purposes as the distribution of reported maize production volumes are right-skewed. We adopt this approach throughout the paper.

[15] Further details are available upon request.

For maize reports we observe some differences in retrospective reporting relative to not including prompts, while for happiness the inclusion of structured prompts does not affect recalled measures. As reported in column (2), providing structured prompts after asking about the 2016 outcome results in a lower 2015 maize recalled value, but the volume produced in 2016 weakens this effect; more specifically, the negative effect of including prompts decreases (and reverses) as the 2016 concurrent report increases. The same occurs for the model in deviations in column (4), where larger positive deviations in the 2016 concurrent report relative to the 2015 concurrent report offset the negative effect of including prompts.

Figure 2 Panel B shows the estimated effects on the 2015 maize recalled value from the inclusion of structured prompts for different percentile values of 2016 concurrent reports. In the case of farmers reporting producing 250 kilograms of maize in 2016 (10 percentile), their 2015 recalled production value is 353 kilograms lower when they are first asked to qualitatively compare their 2016 and 2015 values (relative to not doing such qualitative comparison), while for farmers reporting producing 2,880 kilograms of maize in 2016 (90 percentile), the corresponding 2015 recall values are 702 kilograms higher. For the median 2016 concurrent report of 1,000 kilograms, the estimated effect is negative and close to -52 kilograms, although the effect is not statistically significant.[16]

Similar to reversing the order of questions, adding structured prompts after asking about the most recent outcome helps to reduce mental anchoring when collecting retrospective values. Yet, the reduction in this case is more constrained, compared to the question order reversal case, to the amount produced in the most recent period. The inclusion of structured prompts can be helpful only when the production performance (volume) in the most recent period is low; that is, for 2016 production volumes that are below the 25 percentile. In contrast, for 2016 concurrent reports almost above the 75 percentile, the estimated effect of including prompts, relative to not including them, becomes positive.

4.2.3 Leveraging aggregate (community) data as an anchor

A final survey design approach we test uses past average community data for framing purposes prior to asking about individual outcomes in the previous period. The reasoning behind this survey design is that using the average output reported by the farmer's community in the past period to frame the previous agricultural season could help a farmer to better recall her/his past outcome. This design is only applicable to maize reports and we use the average volume of production reported by the farmer's GAC in 2015 as a frame to elicit 2015 individual outcomes. We separately consider a survey module design that uses the average 2015 GAC value for framing (Treatment 5) and a design that uses both the average 2015 GAC value followed by structured prompts preceding the elicitation of maize production of the household in the previous agricultural season (Treatment 6).

Table 6 shows the estimation results where we compare the retrospective outcomes of these two groups (Treatment 5 and 6) relative to the group that were only asked to directly elicit their 2015 outcome (base Treatment 4). Note that the relevant comparison (base) group in this case is Treatment 4 and not Treatment 1, as this third alternative survey module design does not require eliciting output information for the most recent period (neither prior nor after gathering information for the previous period) as in the

[16] See Appendix Figure B.1 Panel B for the estimated effects of including structured prompts for the model in deviations, which are similar to the model in levels.

cases above.[17] From column (2), we observe that providing the 2015 community reference value (Treatment 5) results in a higher 2015 maize recalled value relative to directly asking about the past production, but this positive shift is attenuated by the volume produced by the farmer in 2016; hence, the positive effect of using an aggregate frame to elicit past outcomes is decreasing in the amount produced in the most recent period and eventually reverses. In contrast, regardless of the amount produced in 2016, providing the 2015 community reference value followed by structured prompts (Treatment 6) does not result in different retrospective outcomes than only asking about the past production. All these behavior patterns for Treatments 5 and 6 (relative to Treatment 4) hold when further controlling for the average 2015 GAC value in the regressions, as reported in column (3). Likewise, the results for the model in deviations presented in columns (5) and (6) show that larger deviations in the 2016 minus 2015 concurrent reports counteract (and reverse) the positive shift of using an aggregate frame (Treatment 5), while using an aggregate frame combined with structured prompts (Treatment 6) does not result in positive (or negative) shifts relative to the base Treatment 4 group, although larger deviations in the 2016 minus 2015 concurrent reports exert a negative effect in this model specification. Note also from column (6) that the larger the deviation of the average 2015 GAC value (used for framing) from the 2015 concurrent value, the larger the deviation of the 2015 recalled value from the 2015 concurrent value.

Figure 3 depicts the estimated effects on the 2015 maize recalled value of using an aggregate frame (Panel A) and using both an aggregate frame and structured prompts (Panel B) for different percentile values of 2016 concurrent reports. In Panel A, for farmers reporting to have produced 250 kilograms of maize in 2016 (10 percentile), their 2015 recalled production volume is about 573 kilograms higher when using an aggregate frame (compared to directly asking about the past volume), while for farmers reporting to have produced 2,880 kilograms of maize in 2016 (90 percentile), the corresponding 2015 recalled volume is, in contrast, 1,060 kilograms lower. For the sample median of 1,000 kilograms produced in 2016, the estimated effect of using a group frame is positive and equivalent to 108 kilograms, but the effect is not significant. We observe a similar pattern of effects in Panel B as the 2016 concurrent report increases, but the effects are not statistically different from zero.[18]

Overall, whether using past average community data aids farmers improving their recall of their past outcome also depends on the amount produced by the farmer in the most recent period. In this case, however, the estimated effect of using an aggregate frame (compared to directly asking about the past outcome) is positive for 2016 production volumes roughly below the 25 percentile and is only negative for volumes above the 85 percentile; that is, for very high production volumes in the most recent period. This approach is hence more likely to rather exacerbate the underlying recall bias problem we are seeking to mitigate. The results of combining an aggregate frame with structured prompts are less clear (precise), irrespective of the amount produced in 2016.

4.3 Extent of bias reduction

To put our results for retrospective maize production in broader perspective, we compare the median conditional deviation of 2015 recalled versus concurrent maize reports for Treatment groups 2, 3, 5, and

[17] Recall that in Treatment 4 module we do not gather output information for the current period.

[18] Appendix Figure B.2 reports the corresponding effects of Treatment 5 and Treatment 6 (relative to Treatment 4) for the model in deviations, which are similar to the effects of the model in levels presented in Figure 3.

6 relative to our base Treatment 1 group. We calculate for each farmer the difference (in absolute terms) between their predicted 2015 recalled value that result from the estimated full-model regressions in levels (reported in Tables 4 through 6) and their observed 2015 concurrent report and obtain the median deviation among each treatment group; we then compare these median conditional deviations across groups using Treatment 1 as the base group.[19] Figure 4 reports the resulting ratios (in percentage) and confidence bands of +/- one standard deviation based on 2,000 bootstrap replications.[20] A ratio below 100% implies an expected bias reduction when implementing one of the alternative survey modules relative to Treatment 1 module, while a ratio above 100% indicates the opposite.

Several important implications emerge from the figure. First, reversing the order of questions in Treatment 2 results in the highest (and more precise) bias reduction among the different survey modules considered to gather self-reported recalled production volumes. Asking first about the past outcome and then to report the most recent outcome entails a 34% bias reduction or lower absolute deviation in the 2015 recalled versus concurrent report, relative to asking first about the present and then about the past (base Treatment 1). Second, using structured prompts prior to asking about the past in Treatment 3 involves a bias reduction, but to a lesser extent (and more imprecise) than Treatment 2; the anticipated bias reduction is 14.5% relative to Treatment 1. Third, using past average community data to frame the previous agricultural season prior to asking about the past (without structured prompts in Treatment 5 and with prompts in Treatment 6) entails, in contrast, a bias increase in the order of 52-54% relative to Treatment 1.[21]

Note that while the focus of the study is on anchoring bias, these estimated ratios may not only represent variations in anchoring bias but could also reflect changes in other confounded recall biases (if applicable) across survey modules. In sum, these findings are in line with the discussion in the previous section and support switching the order of questions (i.e., past before present) when collecting self-reported retrospective production data: while the bias cannot be removed, implementing Treatment 2 survey module anticipates reducing the gap between the recalled and past concurrent report to a larger and more precise extent than other treatment modules.

4.4 Heterogeneity

Lastly, we examine whether the observed mental anchoring in self-reported retrospective measures, documented in Section 4.1, is more salient among specific groups of farmers and, if so, whether any of the different survey module designs considered can be more helpful in reducing the anchoring behavior

[19] Since in Tables 4 and 5 we compare the behavior of farmers in Treatments 2 and 3 against Treatment 1, we can directly derive (and compare) median conditional deviations in recall values for farmers across these three treatment groups. From Table 6, where we compare Treatments 5 and 6 against Treatment 4, we derive median conditional deviations in recall values for farmers in Treatments 5 and 6 and compare them to the median unconditional deviation in recall values of farmers in Treatment 1 (i.e., for the latter we consider the difference in their 2015 reported recall value and their 2015 concurrent report). The results for Treatments 2 and 3 are qualitatively similar if we, alternatively, compare them against the median unconditional deviation of Treatment 1 farmers.

[20] The described exercise is repeated on 2,000 nonparametric bootstrapped samples to derive the corresponding standard deviations.

[21] Relative to Treatment 4, there is neither a bias reduction nor increase when implementing Treatment 5 or 6 modules (the corresponding ratios are 102.4% and 101.4%, respectively).

of these subgroups. Appendix Table B.2 reports the results of exploring systematic differences in mental anchoring among self-reports of maize production (Panel A) and happiness (Panel B) under recall by respondent's sex (if male), age (if age above median), education level (if primary education or higher), household size (if household size above median), land size (if land size above median), and location (if Dowa district). The columns in the table correspond to separate re-estimations of equation (1a) including in each case a heterogenous indicator for these six characteristics (described in parentheses) and the interaction of these indicator with the 2016 concurrent report.[22]

For maize reports, presented in columns (1) through (6), we observe differences by age, household size, and location, while for happiness recalled values we do not find behavior differences across all six characteristics reported in columns (7) through (12). In the case of maize, older respondents, larger households, and farmers located in Dowa district report, on average, lower 2015 recalled values (captured by the heterogeneous indicator variable), but they also exhibit stronger mental anchoring as reflected by the interaction term of the 2016 concurrent report with the heterogenous indicator. From column (2), an additional kilogram of maize reported to have been produced in 2016 is associated with a 0.555-kilogram larger increase in the 2015 recalled volume among farmers older than 42 years old relative to younger farmers. From column (4), one more kilogram of maize produced in 2016 is correlated with a 0.356-kilogram larger increase in the volume recalled for 2015 among households with more than six members relative to households of smaller size, while from column (6) one more kilogram of maize produced in 2016 is correlated with a 0.274-kilogram larger increase in the 2015 recalled volume among farmers located in Dowa district relative to farmers located in Ntchisi district. It follows that older farmers could have more difficulty recalling values than younger ones and thereby exhibit a larger degree of mental anchoring; a similar reasoning could apply to larger households that might experience greater trouble in recalling total production volumes (among several members likely involved in agricultural activities), whereas farmers located in the Dowa district produce relatively smaller volumes of maize compared to other crops that could difficult recalling past maize volumes (Judge and Schechter, 2009).[23]

Based on the larger anchoring behavior for maize reports among older respondents, larger households, and farmers located in Dowa district, in Appendix Table B.3 we further assess whether any of the alternative survey module designs discussed above can be more effective in attenuating the mental anchoring of these three subgroups of farmers. Columns (1) through (3) compare the behavior of each subgroup under Treatment 2 versus 1, columns (4) through (6) compare their behavior under Treatment 3 versus 1, and columns (7) through (9) compare their behavior under Treatments 5 and 6 versus 4. From the reported interactions of the treatment dummies with the heterogeneous indicator variables and the triple interactions of the treatment dummies with the 2016 concurrent reports and heterogeneous indicators, we generally observe that none of the different module treatments is necessarily more helpful to reduce the anchoring behavior of these specific subgroups. The only exception is Treatment 5 (using average community data to frame the past season) for farmers in Dowa district that seem to mitigate the

[22] We focus on the model in levels but the results for the model in deviations are qualitatively similar and available upon request.

[23] Godlonton et al. (2018) also find that larger households show stronger anchoring, compared to smaller households, when recalling their incomes but do not observe systematic differences by respondent's age (neither sex and education).

anchoring role of their 2016 concurrent report when collecting self-reported retrospective data, as reported in column (9).

5. Concluding Remarks

Much of the existing literature on recall bias seeks to quantify the extent of this bias across different settings. While this is important in stressing caution and urging researchers to interpret results with appropriate nuance it is also critical to find avenues to reduce such biases in future work. In this paper, we focus on anchoring bias and test alternative survey design strategies (question order, retrieval cues, and aggregate anchoring without and with retrieval cues) to attenuate mental anchoring in retrospective data collection. We take advantage of a comprehensive longitudinal survey of smallholder farmers in Malawi where we embed our survey design experiment and exploit differences between recall and concurrent responses for both an objective indicator (maize production) and subjective indicator (level of happiness).

Similar to other studies, we first document that respondents consistently use their reported outcome for the most recent period as a cognitive heuristic to aid them recalling their outcome for the previous period. We then assess whether different survey module designs can help reduce these cognitive biases. In the case of maize, we overall find that asking for retrospective data before concurrent data reduces recall bias by approximately 34%, a significant improvement with no increase in data collection time. Retrieval cues (structured prompts) are, in turn, less successful and precise in reducing the bias (14.5%) and more costly in terms of increased data collection time (24 additional seconds per retrospective question that can add up quickly in a lengthy survey fielded to a large population), while aggregate (community) anchors used alone or in conjunction with structured prompts can end up exacerbating the bias. For happiness reports, however, neither reversing the order of questions nor retrieval cues help to alleviate the anchoring bias.

The results thus recommend reversing the order of questions (i.e., asking first about the past and then about the present) when agricultural retrospective data collection is necessary, at least for objective indicators such as volume of production. In particular, a small change in survey design can help to mitigate mental anchoring among smallholders without incurring in extra survey costs. Considering that respondents are prone to cognitive biases when referring to past situations (even for relatively short reference periods) and retrospective data collection is heavily influenced by farmers' most recent situation, this actionable approach can be especially relevant when there is interest in drawing comparisons over self-reported outcomes across different periods that would otherwise result in important mismeasurements.

Lastly, we acknowledge some limitations of the study. First, we focus on a staple crop that is widely produced across the studied area, but the results may not be generalizable to other crops, particularly among crops that comprise a smaller share of households' total production where biases are expected to be larger, and the past before present strategy could be less (or more) effective. This remains an open question for future research. A similar reasoning applies if we are interested in informing survey design for other indicators besides production volume (such as income or working hours) or other contexts and countries, although we hypothesize that smallholder farmers in developing countries are typically exposed to similar volatile patterns in production and labor activities that could largely drive their mental anchoring behavior when recalling past outcomes. In addition, more work is needed to better understand

and address cognitive biases in subjective indicators as the proposed survey designs do not help to ease the anchoring bias in happiness recalled outcomes. Likewise, while we observe larger anchoring behavior in maize reports among older respondents, larger households, and by location, none of the survey module treatments are helpful to lessen mental anchoring among these subgroups. The heterogeneity analysis could be limited though by our relatively small sample size and constitutes another avenue of future work. Finally, while this study shows that asking first about the past and then about the present can help to partially address anchoring behavior when collecting retrospective information for one past period, determining whether a similar solution would apply if we were interested in collecting retrospective information for multiple past periods requires further analysis.

References

Abay, K.A., Abate, G.T., Barrett, C.B., Bernard, T. 2019. Correlated non-classical measurement errors, `Second best' policy inference, and the inverse size-productivity relationship in agriculture. *Journal of Development Economics* 139 (June): 171-184.

Ambler, K., Brauw, A., Godlonton, S. 2018a. Measuring postharvest losses at the farm level in Malawi. *Australian Journal of Agriculture and Resource Economics* 62 (1): 139-60.

Ambler, K., de Brauw, A., Godlonton, S. 2018b. Agricultural support services: Direct effects, complementarities, and time dynamics. IFPRI Discussion Paper 01725.

Ambler, K., Herskowitz, S., Maredia, M. Forthcoming. Are we done yet? Response fatigue and rural livelihoods. *Journal of Development Economics.* https://doi.org/10.1016/j.jdeveco.2021.102736

Aguero, J.M., Frisancho, V. Forthcoming. Measuring violence against women with experimental methods. *Economic Development and Cultural Change*. https://doi.org/10.1086/714008

Arthi, V., Beegle, K., de Weerdt, J., Palacios-Lopez, A. 2016. Not your average job: Irregular schedules, recall bias, and farm labor measurement in Tanzania. *Working Paper*.

Baumeister, R.F., Bratslavsky, E., Finkenauer, C., Vohs, K.D. 2001. Bad is stronger than good. *Review of General Psychology* 5(4): 323-370.

Beegle, K., Carletto, C., Himelein, K. 2012. Reliability of recall in agricultural data. *Journal of Development Economics* 98(1): 34-41.

Beegle, K., de Weerdt, J., Friedman, J., Gibson, J. 2012. Methods of household consumption measurement through surveys: Experimental results from Tanzania. *Journal of Development Economics* 98(1): 3-18.

Beggs, A., Graddy, K. 2009. Anchoring effects: Evidence from art auctions. *American Economic Review* 99(3): 1027-1039.

Campbell, S.D., Sharpe, S.A., 2009. Anchoring bias in consensus forecasts and its effect on market prices. *The Journal of Financial and Quantitative Analysis* 44(2): 369-390.

Deininger, K., C. Carletto, S. Savastano, Muwonge, J. 2012. Can diaries help in improving agricultural production statistics? Evidence from Uganda. *Journal of Development Economics* 98: 42-50.

de Nicola, F., Gine, X. 2014. How accurate is recall data? Evidence from coastal India. *Journal of Development Economics* 106: 55-65.

de Weerdt, J., Beegle, K., Friedman, J., Gibson, J., 2016. The challenge of measuring hunger through survey. *Economic Development and Cultural Change* 64(4): 727-758.

de Weerdt, J., Gibson, J. and Beegle, K. 2020. What can we learn from experimenting with survey methods? *Annual Review of Resource Economics* 12: 431-447.

Frykblom, P., Shogren, J.F. 2000. An experimental testing of anchoring effects in discrete choice questions. *Environmental and Resource Economics* 16: 329-341.

Furnham, A., Chu Boo, H. 2011. A literature review of the anchoring effect. *The Journal of Socio-Economics* 40: 35-42.

Godlonton, S., Hernandez, M.A., Murphy, M. 2018. Anchoring bias in recall data; Evidence from Central America. *American Journal of Agricultural Economics* 100(2): 479-501.

Gollin, D., D. Lagakos, Waugh, M. 2014. The Agricultural Productivity Gap. *Quarterly Journal of Economics* 129(2): 939-993.

Heath, R., Mansuri, G., Rikers, B., Seitz, W., Sharma, D. 2021. Measuring employment: Experimental evidence from Ghana. *The World Bank Economic Review* 35(3): 635-651

Hitczenko, M. 2013. Modeling anchoring effects in sequential likert scale questions. Research Department Working Series No. 13-15, Federal Reserve of Boston.

Hurd, M.D., McFadden, D., Chand, H., Gan, L., Merrill, A., Roberts, M. 1998. Consumption and savings balances of the elderly: Experimental evidence on survey response bias. In D.A. Wise (Ed.), Frontiers in the Economics of Aging (Chapter 8). Massachusetts: University of Chicago Press.

Judge, G., Schecther, L. 2009. Detecting problems in survey data using Benford's Law. *Journal of Human Resources* 44(1): 1-24.

Karlan, D.S., Zinman, J. 2012. List randomization for sensitive behavior: An application for measuring use of loan proceeds. *Journal of Development Economics* 98(1): 71-75.

Krueger, A.B., Schkade, D.A. 2008. The reliability of subjective well-being measures. *Journal of Public Economics* 92(8-9): 1833-1845.

Lobell, D., Azzari, G., Burke, M., Gourlay, S., Jin, Z., Kilic, T., Murray, S. 2019. Eyes in the sky, boots on the ground: Assessing satellite- and ground-based approaches to crop yield measurement and analysis. *American Journal of Agricultural Economic*s 102(1): 209-219.

Loftus, E.F., K.D. Smith, M.R. Klinger, Fiedler, J. 1992. Memory and mismemory for health events. In J.M. Tanur (Ed.), Questions about questions: Inquiries into the cognitive basis of surveys (pp. 102-137). New York: Russel Sage Foundation.

Loftus, E.F., Farthi, D.C. 1985. Retrieving multiple autobiographical memories. *Social Cognition* 3: 280-295.

Schuman, H., Presser, S. 1996. Questions and Answers in Attitude Surveys. Experiments on Question Form, Wording and Context. *SAGE publishing.*

Sudman, S., Bradburn, N.M., Schwarz, N. 1996. Thinking about answers: The application of cognitive processes to survey methodology. San Francisco: Jossey-Bass Publishers.

Tourangeau, R., K. Rasinski, D'Andrade, R. 1991. Attitude structure and belief accessibility. *Journal of Experimental Psychology* 27: 48-75.

Tourangeau, R., Rips, L., Rasinski, K., 2000. The psychology of survey response. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511819322

Tversky, A., Kahneman, D. 1974. Judgement under uncertainty: heuristics and biases. *Science*, 185: 1124-1131.

Walker, W.R., Skowronski, J.J., Thompson, C.P., 2003. Life is pleasant – and memory helps to keep it that way! *Review of General Psychology* 7(2): 203-2010.

Whitten, W.B., Leonard, J.M. 1981. Directed search through autobiographical memory. *Memory and Cognition* 9: 566-579.

Table 1. Survey module treatment designs

Treatment arms	Concurrent measure before retrospective	Structured prompts	Community anchors	Number of observations for maize sample	Number of observations for happiness sample
(T1) Present before Past	Yes	No	No	152	153
(T2) Past before Present	No	No	No	164	168
(T3) T1 + Structured Prompts	Yes	Yes	No	164	167
(T4) Only Past	No^{+}	No	No	168	169
(T5) Framed by Community Anchor	No	No	Yes	142	148
(T6) T5 + Structured Prompts	No	Yes	Yes	151	159
Total				941	964

Note: $^{+}$Treatment 4 module (Only Past) does not include any question about the concurrent (present) value.

Table 2. Summary statistics

Variable	Mean	SD	Min.	Max.
Panel A: Individual and household characteristics at baseline (2014)				
Respondent age	41.809	13.693	17	92
If respondent is male	0.363	0.481	0	1
If respondent has some primary schooling	0.368	0.483	0	1
If respondent has completed primary schooling	0.079	0.269	0	1
If respondent has more than completed primary schooling	0.368	0.483	0	1
Household head age	44.736	13.957	19	92
If household head is male	0.849	0.358	0	1
If polygamous household	0.094	0.292	0	1
Household size	5.587	2.191	1	15
Household land size (in acres)	4.723	3.099	0.500	35.000
If Dowa district	0.529	0.499	0	1
Panel B: Maize production (in kilograms)				
2015 Recalled value	1,133.53	1,429.57	0	27,500
2015 Concurrent report	1,055.64	1,750.40	0	26,000
2016 Concurrent report	1,406.24	2,147.89	0	45,000
2015 Community (GAC) average concurrent report	1,056.65	458.1491	416.804	2,846
Panel C: Happiness level (scale 1-10)				
2015 Recalled value	7.498	3.036	1	10
2015 Concurrent report	8.446	2.309	1	10
2016 Concurrent report	8.784	2.074	1	10
Panel D: Interview characteristics				
If male interviewer in 2014 survey	0.416	0.493	0	1
If male interviewer in 2015 survey	0.521	0.500	0	1
If male interviewer in 2016 survey	0.707	0.455	0	1
Interview length in 2015 survey (in hours)	3.406	1.375	0.218	7.460
Interview length in 2016 survey (in hours)	4.794	2.008	1.250	8.610
Recall module interview length in 2016 survey (in hours)	0.194	0.136	0.060	0.870
Observations				967

Note: The maize production outcomes in Panel B are based on 941 observations (equivalent to the number of households reporting maize production), while the happiness outcomes in Panel C are based on 964 observations (equivalent to the number of households reporting their happiness level). The full working sample includes 967 observations as three households that reported producing maize did not report their happiness level. The 2014 baseline survey was implemented in paper, as opposed to the 2015 and 2016 surveys that were implemented in tablets. During the paper-administered survey interview length was not recorded. SD=Standard deviation.

Table 3. Anchoring behavior in maize and happiness reported outcomes

Coefficient	Maize production (in kilograms)				Happiness level (scale 1-10)			
	Dependent variable:				Dependent variable:			
	2015 Recalled value		2015 Recalled value - 2015 Concurrent report		2015 Recalled value		2015 Recalled value - 2015 Concurrent report	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
2015 Concurrent report	0.162***	0.120**			0.051	0.058		
	(0.057)	(0.052)			(0.045)	(0.045)		
2016 Concurrent report	0.181*	0.169*			0.192***	0.197***		
	(0.101)	(0.090)			(0.051)	(0.051)		
2016 Concurrent report - 2015 Concurrent report			0.415**	0.420**			0.618***	0.616***
			(0.197)	(0.190)			(0.041)	(0.040)
Individual (Household) Characteristics	No	Yes	No	Yes	No	Yes	No	Yes
Interview Characteristics	No	Yes	No	Yes	No	Yes	No	Yes
Observations	941	941	941	941	964	964	964	964

Note: The dependent variable in columns (1), (2), (5) and (6) is the recalled value for 2015 reported in 2016, while the dependent variable in columns (3), (4), (7) and (8) is the difference between the recalled value for 2015 reported in 2016 and the 2015 concurrent report. The 2015 concurrent report is the value reported in 2015 for the outcome in 2015, while the 2016 concurrent report is the value reported in 2016 for the outcome in 2016. Individual (household) characteristics include the age, gender, and level of education of the respondent, age and gender of household head, if household is polygamous, household size, household land size and district location, and were all collected during the 2014 baseline survey. Interview characteristics include the gender of enumerators in all survey rounds, duration of 2015 and 2016 surveys and duration of recall module in 2016 survey. The regressions include indicator variables that take the value of one for missing values on individual and interview characteristics, and zero otherwise; for these cases, we imputed the district median value. The regressions also include indicator variables for surveys partially recorded or paused/re-commenced; for these cases, we assigned the minimum or maximum duration, respectively, of a regular survey in the same district. Robust standard errors reported in parenthesis. Asterisks *, ** and *** denote significance at the 10%, 5% and 1% levels.

Table 4. Past before present (Treatment 2 versus Treatment 1)

Coefficient	Maize production (in kilograms)				Happiness level (scale 1-10)			
	Dependent variable:				Dependent variable:			
	2015 Recalled value		2015 Recalled value - 2015 Concurrent report		2015 Recalled value		2015 Recalled value - 2015 Concurrent report	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
If Treatment 2 [Past before Present]	-116.131	-414.939***	-115.244	-266.355*	0.062	-1.041	0.026	0.006
	(79.774)	(101.471)	(105.572)	(135.395)	(0.346)	(1.596)	(0.393)	(0.399)
If Treatment 2 x 2016 Concurrent report		0.264***				0.125		
		(0.082)				(0.179)		
If Treatment 2 x (2016 Concurrent report - 2015 Concurrent report)				0.724***				0.062
				(0.167)				(0.131)
Concurrent reports	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual (Household) Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Interview Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	316	316	316	316	321	321	321	321

Note: The dependent variable in columns (1), (2), (5) and (6) is the recalled value for 2015 reported in 2016, while the dependent variable in columns (3), (4), (7) and (8) is the difference between the recalled value for 2015 reported in 2016 and the 2015 concurrent report. The 2015 concurrent report is the value reported in 2015 for the outcome in 2015, while the 2016 concurrent report is the value reported in 2016 for the outcome in 2016. The sample corresponds to individuals in Treatment 1 (base category) and Treatment 2 groups. Treatment 1 involves first asking about the 2016 concurrent outcome and then asking to recall the 2015 outcome, while Treatment 2 involves first asking to recall the 2015 outcome and then asking about the 2016 concurrent outcome. Concurrent reports include the 2015 and 2016 concurrent reports in columns (1), (2), (5) and (6), and the difference between the 2016 and 2015 concurrent reports in columns (3), (4), (7) and (8). Individual (household) characteristics include the age, gender, and level of education of the respondent, age and gender of household head, if household is polygamous, household size, household land size and district location, and were all collected during the 2014 baseline survey. Interview characteristics include the gender of enumerators in all survey rounds, duration of 2015 and 2016 surveys and duration of recall module in 2016 survey. The regressions include indicator variables that take the value of one for missing values on individual and interview characteristics, and zero otherwise; for these cases, we imputed the district median value. The regressions also include indicator variables for surveys partially recorded or paused/re-commenced; for these cases, we assigned the minimum or maximum duration, respectively, of a regular survey in the same district. Robust standard errors reported in parenthesis. Asterisks *, ** and *** denote significance at the 10%, 5% and 1% levels.

Table 5. Adding structured prompts (Treatment 3 versus Treatment 1)

Coefficient	Maize production (in kilograms)				Happiness level (scale 1-10)			
	Dependent variable:				Dependent variable:			
	2015 Recalled value		2015 Recalled value - 2015 Concurrent report		2015 Recalled value		2015 Recalled value - 2015 Concurrent report	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
If Treatment 3 [Using Structured Prompts]	94.206	-453.320**	117.019	-147.132	-0.452	-0.631	-0.336	-0.258
	(100.202)	(196.109)	(129.521)	(146.219)	(0.350)	(1.597)	(0.393)	(0.398)
If Treatment 3 x 2016 Concurrent report		0.401***				0.021		
		(0.153)				(0.178)		
If Treatment 3 x (2016 Concurrent report - 2015 Concurrent report)				0.586***				-0.220
				(0.178)				(0.133)
Concurrent reports	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual (Household) Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Interview Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	316	316	316	316	320	320	320	320

Note: The dependent variable in columns (1), (2), (5) and (6) is the recalled value for 2015 reported in 2016, while the dependent variable in columns (3), (4), (7) and (8) is the difference between the recalled value for 2015 reported in 2016 and the 2015 concurrent report. The 2015 concurrent report is the value reported in 2015 for the outcome in 2015, while the 2016 concurrent report is the value reported in 2016 for the outcome in 2016. The sample corresponds to individuals in Treatment 1 (base category) and Treatment 3 groups. Treatment 1 involves first asking about the 2016 concurrent outcome and then asking to recall the 2015 outcome, while Treatment 3 is similar to Treatment 1 but adding structured prompts after asking about the 2016 concurrent outcome and prior to asking to recall the 2015 outcome. Concurrent reports include the 2015 and 2016 concurrent reports in columns (1), (2), (5) and (6), and the difference between the 2016 and 2015 concurrent reports in columns (3), (4), (7) and (8). Individual (household) characteristics include the age, gender, and level of education of the respondent, age and gender of household head, if household is polygamous, household size, household land size and district location, and were all collected during the 2014 baseline survey. Interview characteristics include the gender of enumerators in all survey rounds, duration of 2015 and 2016 surveys and duration of recall module in 2016 survey. The regressions include indicator variables that take the value of one for missing values on individual and interview characteristics, and zero otherwise; for these cases, we imputed the district median value. The regressions also include indicator variables for surveys partially recorded or paused/re-commenced; for these cases, we assigned the minimum or maximum duration, respectively, of a regular survey in the same district. Robust standard errors reported in parenthesis. Asterisks *, ** and *** denote significance at the 10%, 5% and 1% levels.

Table 6. Leveraging aggregate (community) data as an anchor in maize reported outcomes (Treatments 5 and 6 versus Treatment 4)

Coefficient	Maize production (in kilograms)					
	Dependent variable:					
	2015 Recalled value			2015 Recalled value - 2015 Concurrent report		
	(1)	(2)	(3)	(4)	(5)	(6)
If Treatment 5 [Using Community Anchor]	-148.947	728.760**	728.687**	18.698	458.378***	286.821*
	(178.955)	(330.143)	(331.001)	(185.948)	(161.037)	(153.108)
If Treatment 6 [Using Community Anchor + Structured Prompts]	-40.800	258.182	257.868	80.518	134.801	43.537
	(147.352)	(300.588)	(303.999)	(198.737)	(133.231)	(117.260)
If Treatment 5 x 2016 Concurrent report		-0.622**	-0.621**			
		(0.255)	(0.257)			
If Treatment 6 x 2016 Concurrent report		-0.252	-0.252			
		(0.247)	(0.249)			
If Treatment 5 x (2016 Concurrent report - 2015 Concurrent report)					-1.076***	-0.787***
					(0.103)	(0.209)
If Treatment 6 x (2016 Concurrent report - 2015 Concurrent report)					-0.608***	-0.404**
					(0.144)	(0.197)
2015 Group (GAC) average concurrent report			0.004			
			(0.153)			
2015 Group (GAC) average concurrent report - 2015 Concurrent Report						0.300**
						(0.144)
Concurrent reports	Yes	Yes	Yes	Yes	Yes	Yes
Individual (Household) Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Interview Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	461	461	461	461	461	461

Note: The dependent variable in columns (1), (2) and (3) is the recalled value for 2015 reported in 2016, while the dependent variable in columns (4), (5) and (6) is the difference between the recalled value for 2015 reported in 2016 and the 2015 concurrent report. The 2015 concurrent report is the value reported in 2015 for the outcome in 2015, while the 2016 concurrent report is the value reported in 2016 for the outcome in 2016. The sample corresponds to individuals in Treatment 4 (base category), Treatment 5 and Treatment 6 groups. Treatment 4 involves only asking to recall the 2015 outcome, Treatment 5 involves using the 2015 community (GAC) average concurrent report as a framing value prior to asking to recall the 2015 outcome, while Treatment 6 is similar to Treatment 5 but adding structured prompts after using the 2015 community average concurrent report as a framing value and prior to asking to recall the 2015 outcome. Concurrent reports include the 2015 and 2016 concurrent reports in columns (1) through (3) and the difference between the 2016 and 2015 concurrent reports in columns (4) through (6). Individual (household) characteristics include the age, gender, and level of education of the respondent, age and gender of household head, if household is polygamous, household size, household land size and district location, and were all collected during the 2014 baseline survey. Interview characteristics include the gender of enumerators in all survey rounds, duration of 2015 and 2016 surveys and duration of recall module in 2016 survey. The regressions include indicator variables that take the value of one for missing values on individual and interview characteristics, and zero otherwise; for these cases, we imputed the district median value. The regressions also include indicator variables for surveys partially recorded or paused/re-commenced; for these cases, we assigned the minimum or maximum duration, respectively, of a regular survey in the same district. Robust standard errors reported in parenthesis. Asterisks *, ** and *** denote significance at the 10%, 5% and 1% levels.

Figure 1. Data availability by survey year

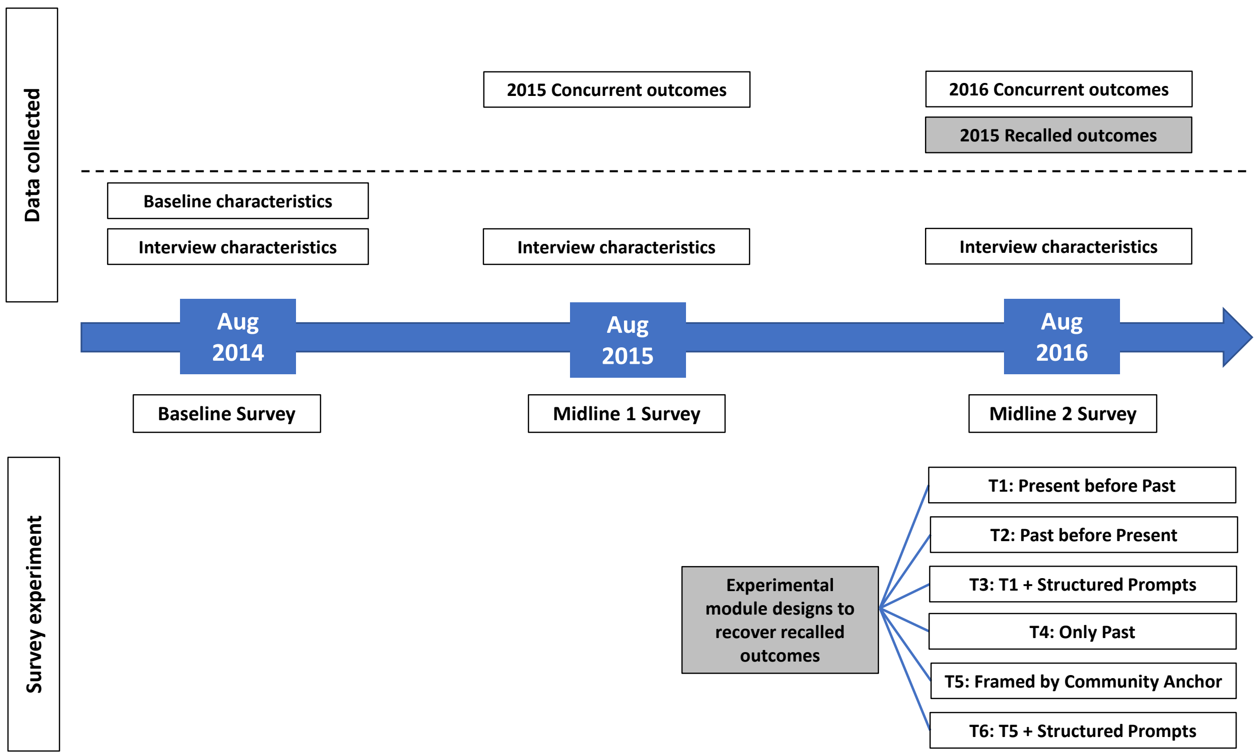
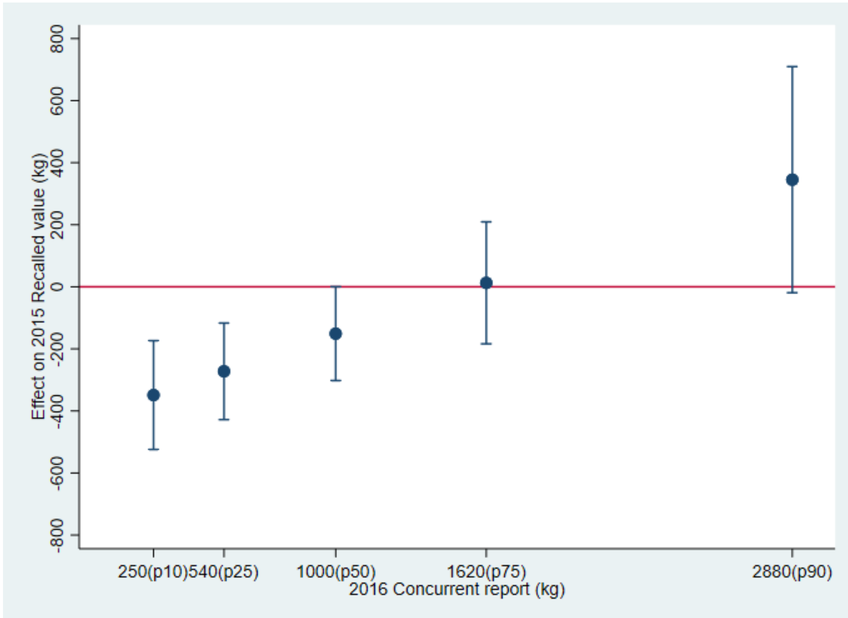
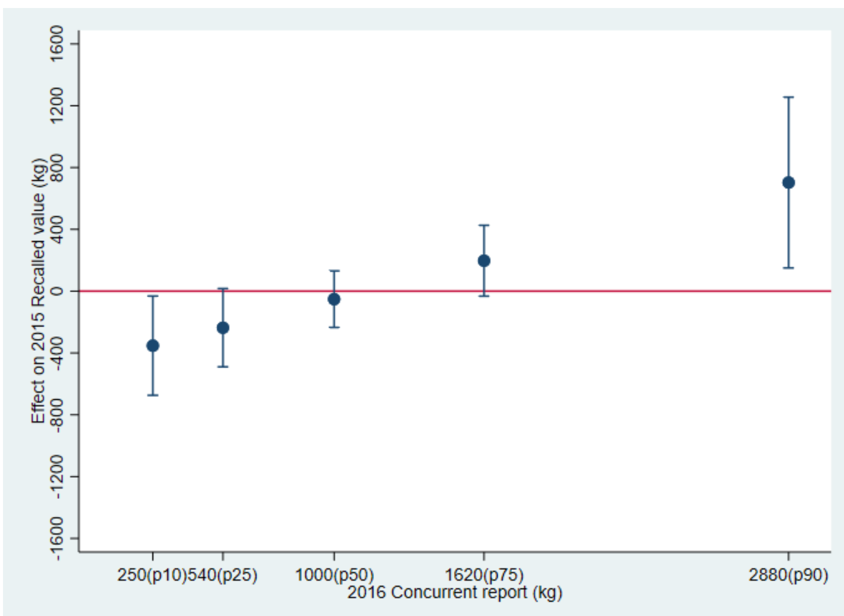


Figure 2. Estimated effects on 2015 maize recalled value for different percentile values of 2016 concurrent reports (Treatments 2 and 3 versus Treatment 1)

Panel A: Past before present (Treatment 2 versus Treatment 1)



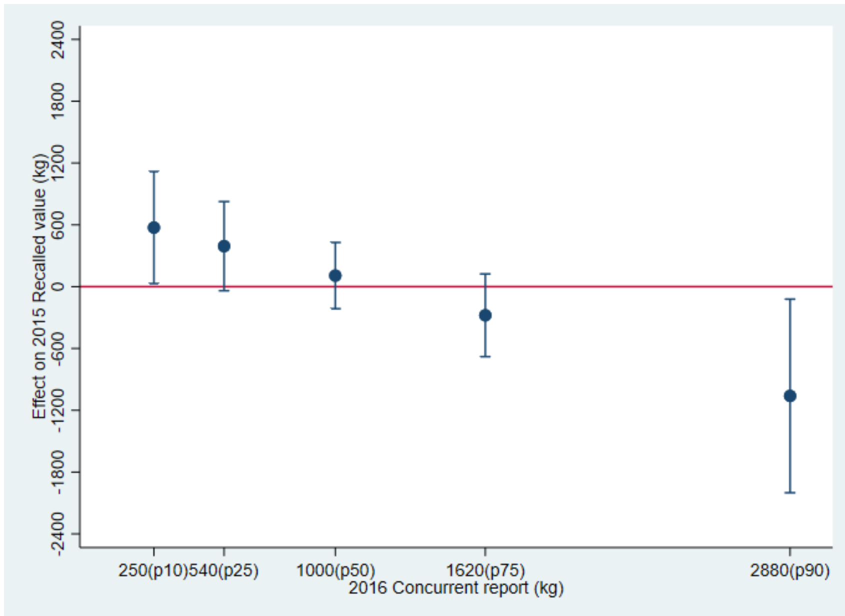
Panel B: Using structured prompts (Treatment 3 versus Treatment 1)



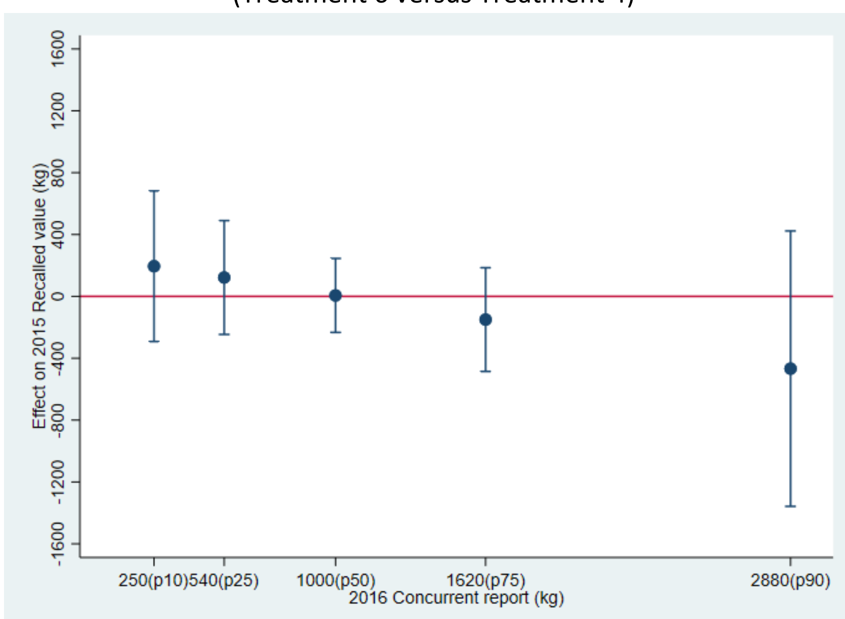
Note: Estimated effects for different percentile values of 2016 maize concurrent reports based on the regression coefficients reported in column (2) of Table 4 (for panel A) and Table 5 (for panel B). The values in the horizontal axis correspond to the 10, 25, 50, 75, and 90 percentiles of the 2016 concurrent reports in the sample. The vertical lines are the 95% confidence intervals of the point estimates derived using the delta method.

Figure 3. Estimated effects on 2015 maize recalled value for different percentile values of 2016 concurrent reports (Treatments 5 and 6 versus Treatment 4)

Panel A: Leveraging community data as anchor
(Treatment 5 versus Treatment 4)

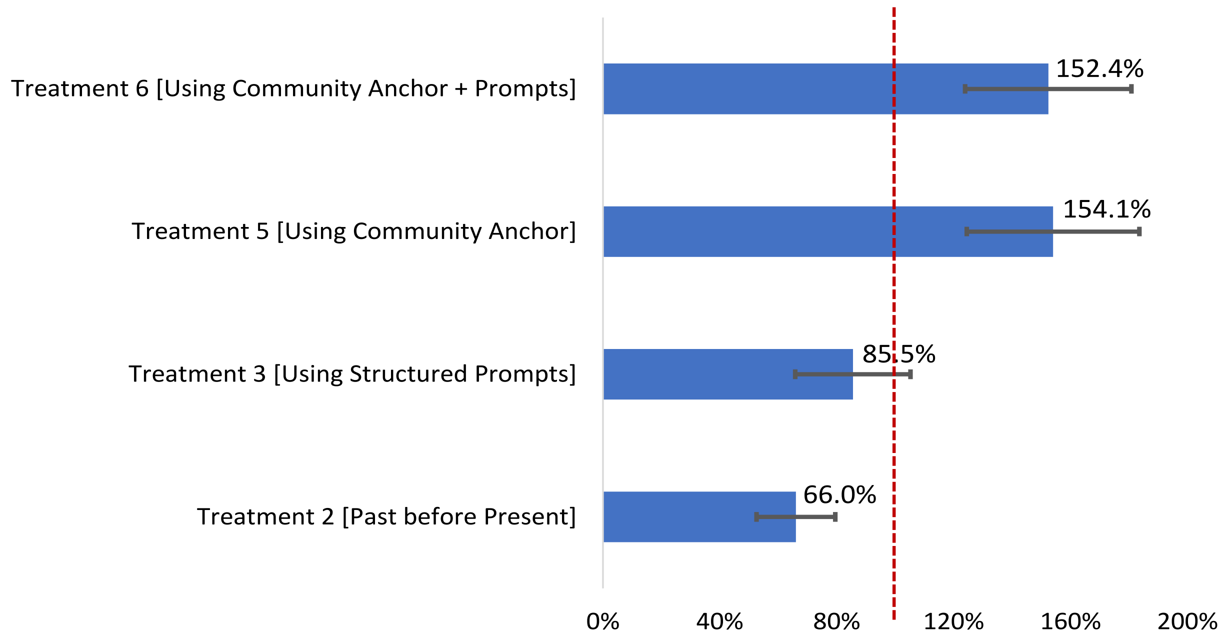


Panel B: Leveraging community data as anchor plus structured prompts
(Treatment 6 versus Treatment 4)



Note: Estimated effects for different percentile values of 2016 maize concurrent reports based on the regression coefficients reported in column (3) of Table 6. The values in the horizontal axis correspond to the 10, 25, 50, 75, and 90 percentiles of the 2016 concurrent reports in the sample. The vertical lines are the 95% confidence intervals of the point estimates derived using the delta method.

Figure 4. Extent of bias reduction in maize reports for different Treatment survey modules, relative to Treatment 1



Note: The base Treatment 1 module involves first asking about the 2016 concurrent outcome and then asking to recall the 2015 outcome (i.e., present before past). The reported ratios (in percentage) correspond to the median conditional deviation of 2015 recalled versus concurrent maize reports (in absolute terms) for Treatment groups 2, 3, 5, and 6 relative to Treatment 1 group. The conditional deviations for each farmer in each group are calculated as the absolute difference between their predicted 2015 recalled value that result from the estimated full-model regressions in levels (reported in Table 4-column (2), Table 5-column (2), and Table 6-column (3)) and their observed 2015 concurrent report. In the case of Treatments 5 and 6 (that are not evaluated against Treatment 1 in Table 6), we compare the derived median conditional deviation for farmers in these groups to the median unconditional deviation among Treatment 1 farmers (i.e., for the latter we consider the difference in their 2015 reported recall values and their 2015 concurrent reports). The horizontal lines are confidence bands of +/- one standard deviation based on 2,000 bootstrap replications.

Appendix A. Survey modules by treatment condition

Treatment Group 1: Present before Past

Maize production	What was the total quantity of production (in kgs) of Maize produced this year?	_______ Kgs
	What was the total quantity of production (in kgs) of Maize produced last year?	_______ Kgs
Happiness	On a scale from 1 to 10 where 1 is the least happy and 10 is the most happy - how happy are you right now?	Number
	On a scale from 1 to 10 where 1 is the least happy and 10 is the most happy - how happy were you this time last year?	Number

Treatment Group 2: Past before Preset

Maize production	What was the total quantity of production (in kgs) of Maize produced last year?	_______ Kgs
	What was the total quantity of production (in kgs) of Maize produced this year?	_______ Kgs
Happiness	On a scale from 1 to 10 where 1 is the least happy and 10 is the most happy - how happy were you this time last year?	Number
	On a scale from 1 to 10 where 1 is the least happy and 10 is the most happy - how happy are you right now?	Number

Treatment Group 3: Present before Past + Structured Prompts

Maize production	What was the total quantity of production (in kgs) of Maize produced this year?	_______ Kgs
	Is this value higher, lower or about the same as last year?	1.Higher 2. Lower 3. About the same
	If higher: Is it much higher, moderately higher, only a little bit higher? If lower: Is it much lower, moderately lower, only a little bit lower?	1. Much higher/lower 2. Moderately higher/lower 3. Only a little bit higher/lower
	What was the total quantity of production (in kgs) of Maize produced last year?	_______ Kgs

Happiness	On a scale from 1 to 10 where 1 is the least happy and 10 is the most happy - how happy are you right now?	Number
	Is this value higher, lower or about the same as last year?	1.Higher 2. Lower 3. About the same
	If higher: Is it much higher, moderately higher, only a little bit higher? If lower: Is it much lower, moderately lower, only a little bit lower?	1. Much higher/lower 2. Moderately higher/lower 3. Only a little bit higher/lower
	On a scale from 1 to 10 where 1 is the least happy and 10 is the most happy - how happy were you this time last year?	Number

Treatment Group 4: Only Past

Maize production	What was the total quantity of production (in kgs) of Maize produced last year?	_______ Kgs

Treatment Group 5: Framed by Community Anchor

Maize production	The average total quantity of production (in kgs) of Maize in your GAC last year was [X]	_______ Kgs
	What was the total quantity of production (in kgs) of Maize produced last year?	_______ Kgs

Treatment Group 6: Framed by Community Anchor + Structured Prompts

Maize production	The average total quantity of production (in kgs) of Maize in your GAC last year was [X]	_______ Kgs
	Do you think the total quantity of production (in kgs) of Maize you produced was higher, lower, or about the same as other farmers in your GAC?	1.Higher 2. Lower 3. About the same
	If higher: Is it much higher, moderately higher, only a little bit higher? If lower: Is it much lower, moderately lower, only a little bit lower?	1. Much higher/lower 2. Moderately higher/lower 3. Only a little bit higher/lower
	What was the total quantity of production (in kgs) of Maize produced last year?	_______ Kgs

Note: GAC is Group Action Committee the farmer belongs to. The modules also included questions for groundnuts and soy (similar to maize), which are not presented as they are not part of this study.

Appendix B. Supplementary Tables and Figures

Table B.1. Omnibus and orthogonality tests across treatment groups

Panel A: Omnibus test (Multinomial logit balance test)		
	Prob > chi2	
	Maize	Happiness
	(1)	(2)
All 6 treatment groups	0.237	0.152
Panel B: Joint orthogonality test across treatment groups		
	p-value	
	Maize	Happiness
	(1)	(2)
Respondent age	0.929	0.832
If respondent is male	0.140	0.106
If respondent has some primary schooling	0.282	0.511
If respondent has completed primary schooling	0.008	0.004
If respondent has more than completed primary schooling	0.125	0.213
Household head age	0.962	0.904
If household head is male	0.237	0.148
If polygamous household	0.791	0.741
Household size	0.098	0.235
Household land size (in acres)	0.932	0.903
If Dowa district	0.129	0.078
If male interviewer in 2014 survey	0.943	0.857
If male interviewer in 2015 survey	0.529	0.523
If male interviewer in 2016 survey	0.611	0.491
Interview length in 2015 survey (in hours)	0.675	0.625
Interview length in 2016 survey (in hours)	0.146	0.112
Recall module interview length in 2016 survey (in hours)	0.020	0.013
Observations	941	964

Note: Panel A reports the probability of rejecting the null hypothesis that the regression coefficients in a multinomial logistic model of the six treatment groups are simultaneously equal to zero; the regressors include individual (household) characteristics collected during 2014 baseline survey (age, gender and level of education of the respondent, age and gender of household head, if household is polygamous, household size, household land size and district location) and interview characteristics (gender of enumerators in all survey rounds, duration of 2015 and 2016 surveys and duration of recall module in 2016 survey). Panel B reports the p-value from the joint orthogonality test across the six treatment groups for each individual (household) and interview characteristic. Robust standard errors are included in the estimations of Panel B.

Table B.2. Heterogeneous anchoring behavior in maize and happiness reported outcomes, model in levels

	Panel A. Maize production (in kilograms)					
	Dependent variable: 2015 Recalled value					
Heterogeneous indicator of interest:	Male respondent	Respondent over 42 years old	Respondent at least completed primary	Household size over 6 members	Land size over 5 acres	Dowa district
	(1)	(2)	(3)	(4)	(5)	(6)
Heterogeneous indicator	-150.385	-745.893***	109.625	-401.026***	296.616	-447.158***
	(260.854)	(279.328)	(222.993)	(134.017)	(253.197)	(134.663)
2015 Concurrent report	0.116**	0.097*	0.119**	0.093**	0.122**	0.101**
	(0.055)	(0.051)	(0.053)	(0.044)	(0.055)	(0.047)
2016 Concurrent report	0.099	0.082*	0.150	0.121	0.149*	0.133
	(0.068)	(0.048)	(0.117)	(0.076)	(0.078)	(0.083)
2016 Concurrent report x	0.204	0.555***	0.053	0.356***	0.032	0.274***
Heterogeneous indicator	(0.184)	(0.151)	(0.151)	(0.106)	(0.154)	(0.103)
Individual (Household) Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Interview Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	941	941	941	941	941	941
	Panel B. Happiness level (scale 1-10)					
	Dependent variable: 2015 Recalled value					
Heterogeneous indicator of interest:	Male respondent	Respondent over 42 years old	Respondent at least completed primary	Household size over 6 members	Land size over 5 acres	Dowa district
	(7)	(8)	(9)	(10)	(11)	(12)
Heterogeneous indicator	-0.855	-0.874	-1.222	1.023	-0.038	-0.920
	(0.926)	(0.946)	(0.911)	(0.914)	(1.026)	(0.907)
2015 Concurrent report	0.055	0.058	0.057	0.058	0.058	0.061
	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)
2016 Concurrent report	0.141**	0.171***	0.161**	0.244***	0.192***	0.158**
	(0.067)	(0.062)	(0.071)	(0.065)	(0.059)	(0.072)
2016 Concurrent report x	0.145	0.056	0.085	-0.126	0.011	0.071
Heterogeneous indicator	(0.103)	(0.104)	(0.101)	(0.102)	(0.114)	(0.100)
Individual Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Interview Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	964	964	964	964	964	964

Note: The dependent variable is the recalled value for 2015 reported in 2016. The 2015 concurrent report is the value reported in 2015 for the outcome in 2015, while the 2016 concurrent report is the value reported in 2016 for the outcome in 2016. The heterogeneous indicator is a dummy variable that varies across columns to identify: male respondents in columns (1) and (7); respondents over 42 years old in columns (2) and (8); respondents that have at least completed primary education in columns (3) and (9); households that have over 6 members in columns (4) and (10); households that have over 5 acres of land in columns (5) and (11); and households located in Dowa district in columns (6) and (12). Individual (household) characteristics include the age, gender, and level of education of the respondent, age and gender of household head, if household is polygamous, household size, household land size and district location, and were all collected during the 2014 baseline survey; in each column, the individual characteristic of interest is replaced by the corresponding heterogeneous indicator. Interview characteristics include the gender of enumerators in all survey rounds, duration of 2015 and 2016 surveys and duration of recall module in 2016 survey. The regressions include indicator variables that take the value of one for missing values on individual and interview characteristics, and zero otherwise; for these cases, we imputed the district median value. The regressions also include indicator variables for surveys partially recorded or paused/re-commenced; for these cases, we assigned the minimum or maximum duration, respectively, of a regular survey in the same district. Robust standard errors reported in parenthesis. Asterisks *, ** and *** denote significance at the 10%, 5% and 1% levels.

Table B.3. Heterogeneous anchoring behavior in maize reported outcomes and alternative module designs, model in levels

Coefficient	Maize production (in kilograms)								
	Dependent variable: 2015 Recalled value								
	Treatment 2 versus Treatment 1			Treatment 3 versus Treatment 1			Treatments 5 and 6 versus Treatment 4		
	Respondent over 42 years old	Household size over 6 members	Dowa district	Respondent over 42 years old	Household size over 6 members	Dowa district	Respondent over 42 years old	Household size over 6 members	Dowa district
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Heterogeneous indicator	-265.977	-349.148**	-603.583***	-494.344**	-400.351**	-429.399**	-1,133.578**	-662.683**	-354.511
	(172.672)	(162.206)	(170.156)	(198.686)	(173.504)	(198.534)	(532.729)	(283.294)	(287.521)
If Treatment 2 [Past before Present]	-73.090	40.583	-36.975						
x Heterogeneous indicator	(149.930)	(176.330)	(156.761)						
If Treatment 3 [Using Structured Prompts]				-92.933	-324.925	72.106			
x Heterogeneous indicator				(224.187)	(301.960)	(214.535)			
If Treatment 5 [Using Community Anchor]							407.159	610.298	134.431
x Heterogeneous indicator							(405.842)	(383.584)	(287.324)
If Treatment 6 [Using Community Anchor +							345.125	255.577	-142.330
Structured Prompts] x Heterogeneous indicator							(427.697)	(265.254)	(242.911)
2016 Concurrent report x Heterogeneous	0.374***	0.424***	0.253	0.446***	0.422***	0.215	0.638*	0.396**	0.463***
indicator	(0.101)	(0.102)	(0.154)	(0.111)	(0.105)	(0.180)	(0.364)	(0.183)	(0.171)
If Treatment 2 x 2016 Concurrent report x	-0.094	-0.188	0.005						
Heterogeneous indicator	(0.106)	(0.114)	(0.172)						
If Treatment 3 x 2016 Concurrent report x				0.131	0.110	0.072			
Heterogeneous indicator				(0.158)	(0.199)	(0.219)			
If Treatment 5 x 2016 Concurrent report x							-0.214	-0.261	-0.410**
Heterogeneous indicator							(0.390)	(0.241)	(0.202)
If Treatment 6 x 2016 Concurrent report x							-0.161	-0.050	-0.103
Heterogeneous indicator							(0.361)	(0.173)	(0.154)

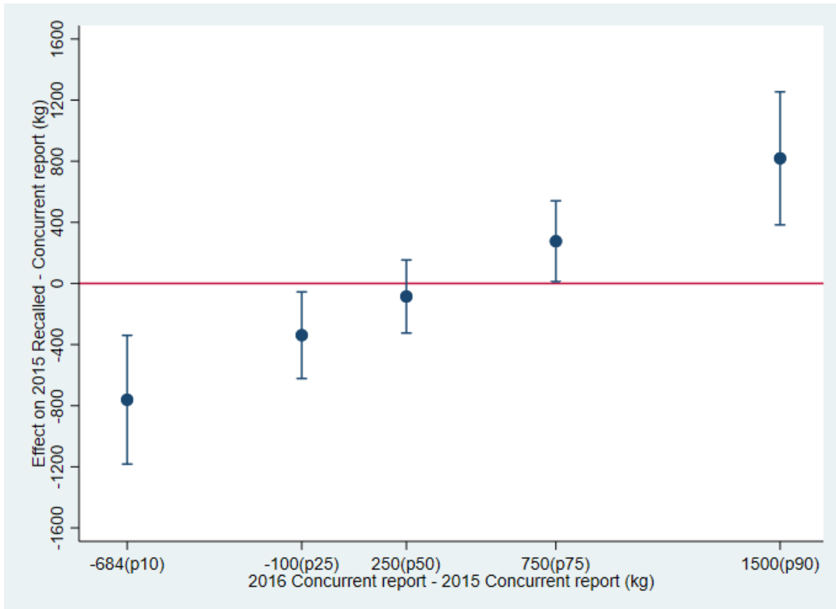
(*Cont.*)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Concurrent reports	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual (Household) Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Interview Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	316	316	316	316	316	316	461	461	461

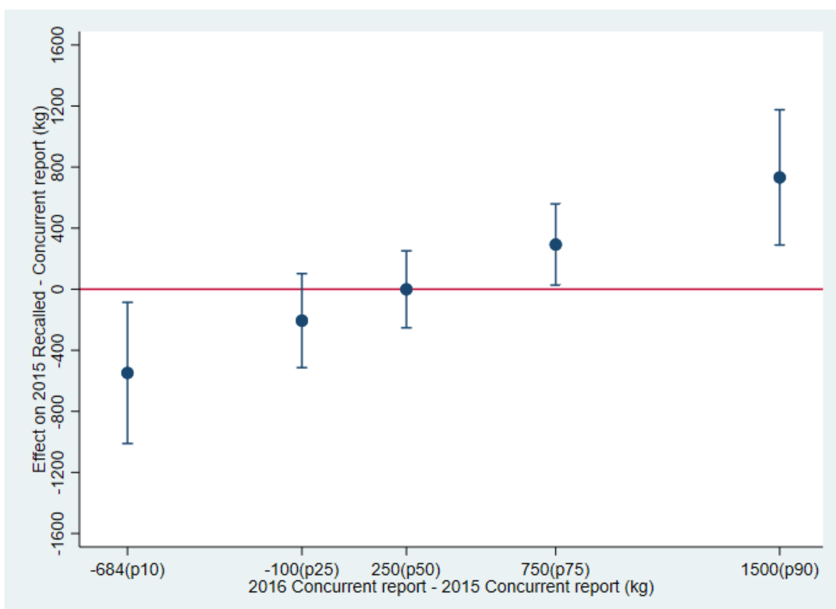
Note: The dependent variable is the recalled value for 2015 reported in 2016. The sample in columns (1) through (3) corresponds to individuals in Treatment 1 (base category) and Treatment 2 groups; Treatment 1 involves first asking about the 2016 concurrent outcome and then asking to recall the 2015 outcome, while Treatment 2 involves first asking to recall the 2015 outcome and then asking about the 2016 concurrent outcome. The sample in columns (4) through (6) corresponds to individuals in Treatment 1 (base category) and Treatment 3 groups; Treatment 3 is similar to Treatment 1 but adding structured prompts after asking about the 2016 concurrent outcome and prior to asking to recall the 2015 outcome. The sample in columns (7) through (9) corresponds to individuals in Treatment 4 (base category), Treatment 5 and Treatment 6 groups; Treatment 4 involves only asking to recall the 2015 outcome, Treatment 5 involves using the 2015 community (GAC) average concurrent report as a framing value prior to asking to recall the 2015 outcome, while Treatment 6 is similar to Treatment 5 but adding structured prompts after using the 2015 community average concurrent report as a framing value and prior to asking to recall the 2015 outcome. The heterogeneous indicator is a dummy variable that varies across columns to identify: respondents over 42 years old in columns (1), (4) and (7); households that have over 6 members in columns (2), (5) and (8); and households located in Dowa district in columns (3), (6) and (9). Concurrent reports include the value reported in 2015 for the outcome in 2015 and the value reported in 2016 for the outcome in 2016; columns (7) through (9) further include the 2015 group (GAC) average concurrent report. Individual (household) characteristics include the age, gender, and level of education of the respondent, age and gender of household head, if household is polygamous, household size, household land size and district location, and were all collected during the 2014 baseline survey; in each column, the individual characteristic of interest is replaced by the corresponding heterogeneous indicator. Interview characteristics include the gender of enumerators in all survey rounds, duration of 2015 and 2016 surveys and duration of recall module in 2016 survey. The regressions include indicator variables that take the value of one for missing values on individual and interview characteristics, and zero otherwise; for these cases, we imputed the district median value. The regressions also include indicator variables for surveys partially recorded or paused/re-commenced; for these cases, we assigned the minimum or maximum duration, respectively, of a regular survey in the same district. Robust standard errors reported in parenthesis. Asterisks *, ** and *** denote significance at the 10%, 5% and 1% levels.

Figure B.1. Estimated effects on 2015 maize recalled value for different percentile values of 2016 concurrent reports, in deviations from 2015 concurrent report (Treatments 2 and 3 versus Treatment 1)

Panel A: Past before present (Treatment 2 versus Treatment 1)



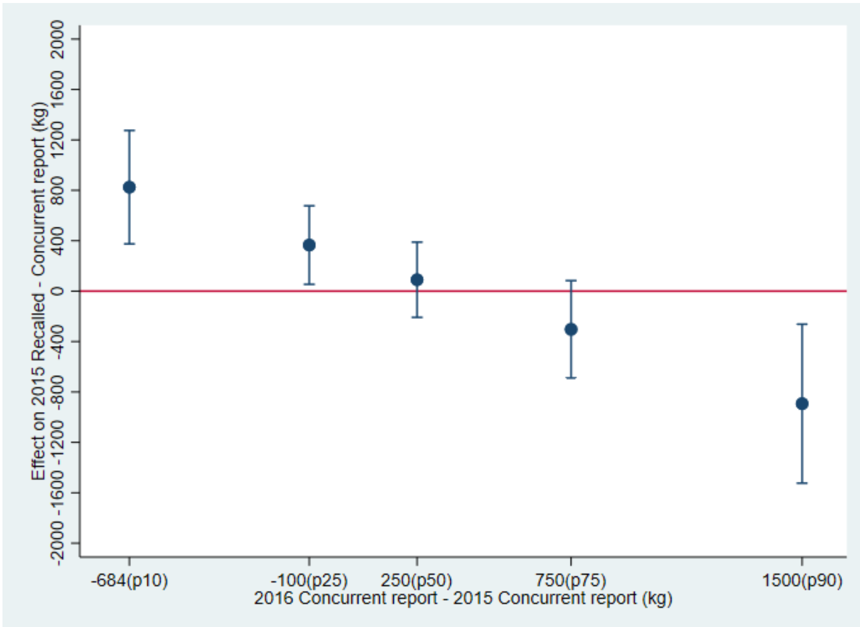
Panel B: Using structured prompts (Treatment 3 versus Treatment 1)



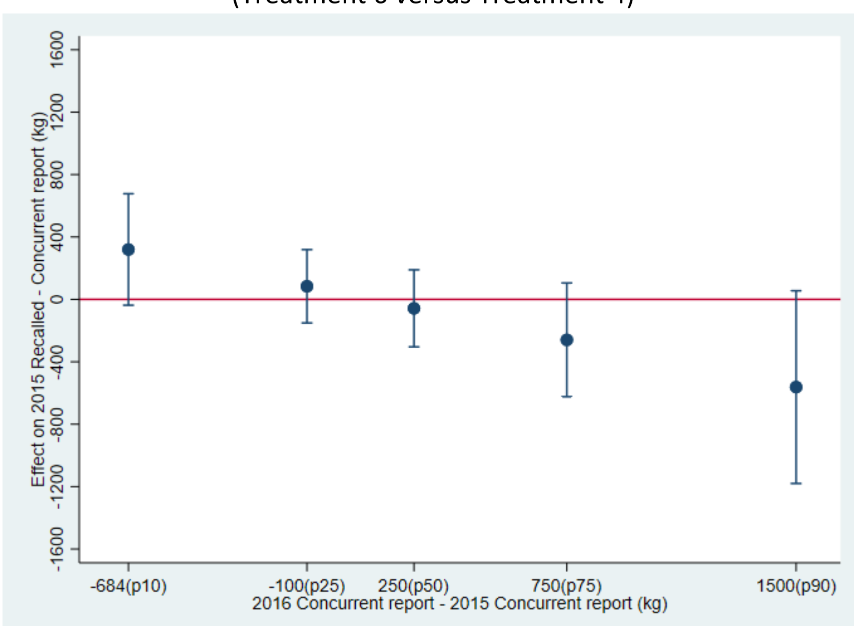
Note: Estimated effects for different percentile values of the difference between 2016 and 2015 maize concurrent reports based on the regression coefficients reported in column (4) of Table 4 (for panel A) and Table 5 (for panel B). The values in the horizontal axis correspond to the 10, 25, 50, 75, and 90 percentiles of the difference between the 2016 and 2015 concurrent reports in the sample. The vertical lines are the 95% confidence intervals of the point estimates derived using the delta method.

Figure B.2. Estimated effects on 2015 maize recalled value for different percentile values of 2016 concurrent reports, in deviations from 2015 concurrent report (Treatments 5 and 6 versus Treatment 4)

Panel A: Leveraging community data as anchor
(Treatment 5 versus Treatment 4)



Panel B: Leveraging community data as anchor plus structured prompts
(Treatment 6 versus Treatment 4)



Note: Estimated effects for different percentile values of the difference between 2016 and 2015 maize concurrent reports based on the regression coefficients reported in column (6) of Table 6. The values in the horizontal axis correspond to the 10, 25, 50, 75, and 90 percentiles of the difference between the 2016 and 2015 concurrent reports in the sample. The vertical lines are the 95% confidence intervals of the point estimates derived using the delta method.