Received: 1 October 2023 | Accepted: 27 August 2024

DOI: 10.1111/dpr.12813

ARTICLE

WILEY

Improving survey quality using paradata: Lessons from a field survey in India

Deepti Goel[1] | **Rosa Abraham**[2]

[1]Pitzer College, USA

[2]Azim Premji University, India

Correspondence
Deepti Goel, Pitzer College, USA.
Email: deepti_goel@pitzer.edu

Funding information
National Council for Applied Economics Research; Indian Institute of Management Bangalore; Institute for What Works to Advance Gender Equality; Azim Premji University

Abstract

Motivation: When collecting evidence from the field, the quality of the data determines the reliability of the analysis. When data are collected in the field by enumerators, the latter's performance needs to be monitored to avoid errant behaviour that could compromise data quality.

Purpose: We show how paradata on the process of data collection itself can improve enumerator performance, using a household survey in India as a case study.

Approach and methods: We conducted action research to improve data quality in the India Working Study conducted in early 2020 in Karnataka and Rajasthan. We designed indicators (flags) from the paradata to mark potential deviant enumerator behaviour in the early stages of the survey. Flagged enumerators were contacted by supervisors who provided constructive feedback. We then measured the performance of the flagged enumerators over the remainder of the survey.

We were able to benchmark specific groups of enumerators facing similar field conditions, namely location and gender of respondents. This allowed us to compare enumerators to a subset of their peers, rather than the entire set of enumerators.

Findings: Our feedback improved enumerator behaviour in the field: flagged enumerators subsequently spent more time on a core module of the questionnaire.

Policy implications: In any survey, two objectives compete: completing a fixed number of interviews per day to reduce costs versus enumerators spending enough time with each respondent to collect meaningful data. To strike a balance between these competing demands, we recommend tracking three paradata indicators: count of completed interviews; average time per completed interview; and ratio of completed to initiated interviews.

We recommend using paradata to improve the quality of data when surveying, thereby reducing standard errors for estimates based on the data and leading to more reliable analysis.

KEYWORDS
data quality, household surveys, India, interviewer effects, paradata

1 | INTRODUCTION

Given the sheer volume of critical tasks to be completed before the start of any survey, designing effective quality control for survey implementation does not always receive the attention it deserves. This is especially true when a small team of individual researchers is in charge of managing the survey, a phenomenon that has gained traction in many low- and middle-income countries (LMICs) (Lupu & Michelitch, 2018). Compromising on quality control is also more likely when compounded by budget, time, and skilled workforce constraints, more commonplace in LMICs. Furthermore, there is evidence suggesting the existence of higher survey fraud in these countries (Kuriakose & Robbins, 2018). All this is worrying, as the efficacy of survey-based policy recommendations depends on the quality of data collected. Against this backdrop, we provide a detailed description of how we used paradata to improve the quality of the India Working Survey (IWS), a field-based household survey implemented in two Indian states in 2020.[1]

Paradata refers to data about the process of data collection (Couper, 1998). They usually include data on who conducted the interview, time stamps for the start and end of the full interview and for individual sections and questions, and revisit information. They could also include keystrokes in case of computer-aided personal interviewing (CAPI), global positioning system (GPS) co-ordinates of enumerator movement and interview location, interviewer characteristics, interviewer observations, and audio/video recordings of respondent–interviewer interaction.

The first wave of IWS was conducted via CAPI, and here we describe how paradata from this wave was used to design and implement a method aimed at streamlining enumerator field practices to reduce interviewer-induced measurement error, an important component of "total survey error" (Olson et al., 2020). We show that our method was effective in changing enumerator behaviour in the field.

A rich body of work exists on post-survey use of paradata to assess and correct for non-response error (Ackermann-Piek et al., 2020; Krueger & West, 2014; Kunz et al., 2024; Pashazadeh et al., 2020) and measurement error (Da Silva & Skinner, 2020; Yan & Olson, 2013). In contrast, work on paradata use concurrent with survey implementation is still emerging, and we add to this literature: Edwards et al. (2017), Edwards et al. (2020), and Mohadjer & Edwards (2018) used anomalies in the data to provide immediate feedback to interviewers as data were being collected. Our exercise is similar, but we innovate on the method to detect anomalies.

Our method is embedded within the statistical process control perspective which advocates adopting procedures used in quality control of industrial products to improve ongoing surveys (Kreuter et al., 2010). Hood and Bushery (1997) and Bushery et al. (1999) are early studies in this tradition. They identified outlier interviews in the U.S. Census Bureau household surveys, interviews which were then redone. In more recent work, Kosyakova et al. (2019) used statistical techniques to identify a confirmed fraudulent interviewer in survey data from Germany.

These studies test the effectiveness of fraud detection methods after data collection is over, whereas we identify and correct deviant behaviour while the survey is ongoing. Moreover, the way we benchmark processes to identify deviant behaviour differs from what others have done.[2] Our method is very close in spirit to that of Guyer et al. (2021), who designed a paradata-driven tool for managing interviewer performance in real time. However, a crucial methodological aspect that we emphasize, namely "dynamic benchmarking" within a group of enumerators facing similar external environments, receives only a passing mention in their study. Our method also has conceptual parallels with the propensity-adjusted interviewer performance (PAIP) indicator developed by West and Groves (2013). They used paradata to predict the probability of completing an interview and evaluated interviewers based on mean deviation between "actual completion" and "predicted probability of completion."

[1]IWS has seven principal investigators (PIs) including the authors of this article. "We," can refer to either all the PIs or only the authors.

[2]Hood and Bushery (1997) used covariates from a previous census for benchmarking, Bushery et al. (1999) used historical survey averages, and Kosyakova et al. (2019) used averages across all interviews.

The role of predicted probability in their method is similar to that of "comparison group's average" in ours.[3] However, they did not update their prediction model over time, while we rely on "dynamic benchmarking" to improve comparison groups' averages so that our averages incorporate information from newer data that becomes available as the survey progresses.

Barring some recent work (Bhuiyan & Lackie, 2016; Choumert-Nkolo et al., 2019; Finn & Ranchhod, 2015), most illustrations of paradata use (including those cited above) are from high-income countries. Although the underlying statistical theory is portable across contexts, operational issues in LMICs are very different (Lupu & Michelitch, 2018). Drawing from our experience, we highlight the difficulties of working with paradata in a resource-constrained context. An important insight from our experience is that established best practices, most of which have evolved from high-income-country experiences, may need to be modified to suit local contexts.

2 | INDIA WORKING SURVEY

The IWS was conducted in Karnataka and Rajasthan with the aim of understanding how social identities, specifically caste, gender, and religion, influence livelihood outcomes. Data collection for the first IWS wave was outsourced to a private agency and was scheduled from February through April 2020. However, field operations had to be stopped in mid-March due to COVID-19.

The agency's enumerators contacted 6,900 respondents from 3,623 households between February 3 and March 17, 2020. Whenever present, one adult female and one adult male from each household were selected at random to be interviewed. Given the gender-sensitive nature of some questions, female respondents were only interviewed by female enumerators and likewise for males. Table 1 shows the organization of the IWS questionnaire according to section. The columns labelled "Female" and "Male" show the approximate count of questions in each section fielded by the female and male enumerators, respectively. As seen, the length of the questionnaire varied according to the gender of the enumerator or the respondent.

Every couple of days throughout the survey period, we received two files from the agency that contained the survey data and paradata collected until then. In both files there is a one-to-one correspondence between an observation/row and a respondent. We therefore use the terms observation, respondent, and interview interchangeably. Table 2 presents an exhaustive list of paradata variables used in this article. These were captured via CAPI at the time of administering the questionnaire.

3 | FLAGGING DEVIANT ENUMERATORS

We used "flags" to identify enumerators who exhibit deviant practices. Although the basic idea of a flag (explained below) is not novel (Jans et al., 2013), its use to improve data quality while a survey is in progress is not customary as yet.

A flag is basically a warning that marks deviation from expected practice. It involves comparing within a group of enumerators who face similar field conditions and marking those enumerators, if any, whose performance deviated substantially from the group's average. We call each such enumerator group a "comparison group." Restricting comparisons to enumerators facing similar conditions allows us to credibly interpret the group average as the process's steady state, so deviations from this average indicate possible errant behaviour requiring

[3]We identify a deviant practice as a large difference between an enumerator's actual performance and the average performance in their comparison group. We discuss details in Section 3.

TABLE 1 IWS Questionnaire (main sections and number of questions by respondent gender).

Section	Description	Female	Male
0. Household Register	Basic household roster. Fielded only by female enumerators.	6	
1. Demographic Characteristics	Demographic information such as caste, education, and major work status. Female enumerators recorded it for all household members, while male enumerators for only the male respondent.	22	17
2. Household Living Standards	Dwelling information, household amenities, and assets. Fielded only by female enumerators.	12	
3. Activity Profile for the Last Year	Major work activity status and skill set of the respondent.	26	26
4. Weekly Labour Force Status	Details about respondent's work activities in the week prior to the interview. Core section of IWS.	58	58
5. Household Production Activities	Time spent by the respondent on household production activities the day before the interview.	12	12
6. Life History Calendar	This section was administered on paper. Paradata is not available for it.		
7. Discrimination	Attitudes regarding gender, caste and religion in relation to livelihood and experiences of discrimination at work.	30	30
8. Decision Making	How are decisions made within the household?	12	12
9. Intergenerational Mobility	Respondent's parents' education and occupation.	9	9
10. Social Networks	Respondent's social contacts and the help received from them.	5	5
11. Women Out of Work Force	Information about women who reported 'not working' as their major work status.	9	9
12. Students	Information about respondents who reported 'studying or attending an education institution' as their major work status.	2	2
13. Unemployed	Information about respondents who reported being 'unemployed' as their major work status.	9	9

intervention. A flag, however, is only suggestive of faulty behaviour. It is possible, though unlikely,[4] that the deviant behaviour was the right response under the circumstances faced by the enumerator. Field supervisors should therefore refrain from accusing flagged enumerators of outright malpractice.

Next, we discuss crucial design features of our method for creating paradata-based flags.

3.1 | Defining a comparison group

Optimally defined comparison groups maximize between-group variability and minimize within-group variability under stable field conditions. This ensures that multiple data-generating processes do not operate within the same comparison group. Mixing different processes is flawed because, depending on which process is driving a given observation, an associated field practice may be justifiably amplified or attenuated relative to the group's average

[4]It is unlikely because when flagging deviant behaviour we are only comparing among enumerators who faced similar field conditions. We therefore expect them to behave similarly.

TABLE 2 IWS Paradata Variables.

Variable	Description
enumerator.id	Unique identifier associated with each enumerator.
enumerator.gender	Gender (male/female) of the enumerator.
respondent.id	Unique identifier associated with each respondent.
interview.id	Same as respondent.id.
state	State (Karnataka/Rajasthan) of the respondent.
region	Region of residence (rural/urban) of the respondent.
consent	Whether or not the respondent consented to the interview.
interview.start.stamp	Date (dd-mm-yyyy) and time (hrs: mins) when interview started.
interview.end.stamp	Date (dd-mm-yyyy) and time (hrs: mins) when interview ended. Incomplete interviews also have an end stamp.
interview.duration	Time between start and end of the interview. Only includes the time that the enumerator spent with the respondent administering the survey questions. If the interview was conducted in multiple spells, it does not include the time between spells.
section.duration#	Time between the start and end of each section of the questionnaire. There is one such variable for each section.
revisits	Number of additional visits made to interview the respondent.
visit.result	The final completion status of the interview at the time of ending it. This is as marked by the enumerator

and it would be incorrect to term it errant behaviour. In IWS, we defined a comparison group as a specific state (Karnataka or Rajasthan), subregion (urban or rural), and the enumerator's gender (male or female) combination, resulting in eight such groups. Next, we explain the rationale for including each delineating dimension.

Consider enumerators operating in the same state and subregion. Recall that a respondent had to be interviewed by an enumerator of the same gender. It would therefore be incorrect to bracket male and female enumerators, as the data-generating process for each enumerator type differs by the gender of their respondents—we already anticipate them to follow different protocols because the IWS questionnaire is longer for female enumerators, entailing longer interview times. Putting male and female enumerators in the same group would increase within-group variability, violating a defining feature of a comparison group. Following similar reasoning, we differentiate by state and subregion.

Lastly, to expect enumerators within a comparison group to exhibit similar behaviour on average, we must establish that there was no systematic matching of respondents with enumerators within these groups. To make this claim, we present some details about IWS sampling. Field supervisors split the enumerators under their charge into pairs consisting of one male and one female enumerator, who remained together until an entire village (block)[5] was surveyed. Supervisors randomly assigned households to each enumerator pair from among households that had been previously selected at random from the village (block) by personnel not involved in the enumeration. An enumerator pair visited each assigned household and created the household roster whenever consent was given. From the roster, an adult male and female were once again randomly selected to serve as respondents. Typically, the enumerator pair that initiated contact with a household also carried out all the follow-up interviews. Since a comparison group consists of several villages (blocks), we must establish that there was no systematic assignment of enumerators across villages (blocks). We rule this out because the personnel who allocated enumerators across villages (blocks) did not have adequate information

[5]A village (block) is a geographic cluster of households in a rural (urban) setting.

about selected villages (blocks) to be able to carry out any kind of matching; villages (blocks) were randomly selected by the principal investigators (PIs) from census frames.

3.2 | Setting performance window length

A highlight of our method is "dynamic benchmarking." Instead of examining enumerator performance cumulatively, we studied it in blocks of a week at a time. Consequently, each practice was evaluated against a moving average, taking into account all changes over time. For instance, enumerators predictably gain proficiency as the survey progresses, and separate performance windows account for this in terms of progressively shorter average interview times.

There is a trade-off when deciding the appropriate window length. Too long a window would imply that faulty practices continue unchecked. On the other hand, if it is too short, there may not be enough data points within comparison groups for the underlying statistical theory to generate credible flags. Additionally, shorter windows dictate more frequent interventions, taking up supervisors' time. In retrospect, for IWS a two-week window instead of a weekly one may have been better at managing this trade-off.

3.3 | Setting thresholds for errant behaviour

How far away from the group mean should a value be for it to be flagged? Some studies adopt the three-sigma rule, namely, three standard deviations away from the mean, as a statistical benchmark (Jans et al., 2013). The authors of such studies acknowledge that no single rule fits all.

We used two somewhat arbitrary thresholds of 1 and 1.6 standard deviations from the mean. We believe that the best way to set thresholds is to base them on a pilot phase. Pilot data would give a sense of the underlying variance in each practice that can then be used to set meaningful practice-specific thresholds.[6]

3.4 | Choosing flags

Each flag is associated with a specific field practice worth monitoring. Table 3 presents the IWS flags. In column (3), against each flag, we specify the main performance dimension(s) it evaluates. We considered three dimensions: (a) "Content knowledge," which refers to a sound understanding of the concepts used in the questionnaire; (b) "Effort exerted" refers to an enumerator's effort proxied by time spent with the respondent; and (c) "Ethics," which is about adherence to prescribed interview protocols. In columns (4) and (5) we describe each flag in terms of the field practice it monitors and the underlying concern raised by outliers, respectively. The last three columns provide details about flag design and cover: (a) whether the flag was constructed using paradata or the main survey data; (b) the criteria used for flagging interviews/enumerators and the corresponding thresholds wherever applicable; (c) the rule used for identifying enumerators for intervention. Our choice of flags is neither prescriptive nor exhaustive. In choosing these flags, we were driven by a focus on collecting high quality data and not so much by survey timelines; the data collection agency had enough checks against overshooting timelines, and we were more concerned about enumerators violating protocols to meet the agency's productivity targets.

We caution against creating too many flags, even though the marginal cost of designing a flag is small. Having too many flags means crucial feedback to enumerators may be lost in translation; especially if there are nuanced

[6]We could not implement paradata piloting because of tight timelines for rolling out the survey.

TABLE 3 IWS Flags to Identify Deviant Behaviour.

S. no.	Flag name	Dimension being evaluated	Field practice being monitored	Underlying concern	Based on paradata or survey data	Threshold	Ranking enumerators for intervention
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
1	Survey Time	Effort	Time taken to field select sections.[1]	Very short interview time suggests violation of interview protocols.	Paradata	Interview got flagged if its standardized survey time[2] is below −1.6, OR its raw survey time is less than 10 minutes.	For each enumerator, ratio of flagged interviews to completed interviews in that week was calculated. Within each state-region-gender strata, top three enumerators with highest positive ratios (if any) were flagged for intervention.
2	Section 0 Time	Effort	Time taken to field 'Household Register'.	Very short section time suggests violation of interview protocols.	Paradata	Interview got flagged if its standardized section time[2] is below −1.6.	For each enumerator, ratio of flagged interviews to completed interviews in that week was calculated. Within each state-region-gender strata, top three enumerators with highest positive ratios (if any) were flagged for intervention.
3	Section 1 Time	Effort	Time taken to field 'Demographic Characteristics'.	Very short section time suggests violation of interview protocols.	Paradata	Interview got flagged if its standardized section time[2] is below −1.6.	For each enumerator, ratio of flagged interviews to completed interviews in that week was calculated. Within each state-region-gender strata, top three enumerators with highest positive ratios (if any) were flagged for intervention.

(Continues)

TABLE 3 (Continued)

S. no.	Flag name	Dimension being evaluated	Field practice being monitored	Underlying concern	Based on paradata or survey data	Threshold	Ranking enumerators for intervention
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
4	Section 2 Time	Effort	Time taken to field 'Household Living Standards'.	Very short section time suggests violation of interview protocols.	Paradata	Interview got flagged if its standardized section time^2 is below −1.6.	For each enumerator, ratio of flagged interviews to completed interviews in that week was calculated. Within each state-region-gender strata, top three enumerators with highest positive ratios (if any) were flagged for intervention.
5	Section 3 Time	Effort	Time taken to field 'Activity Profile for the Last Year'.	Very short section time suggests violation of interview protocols.	Paradata	Interview got flagged if its standardized section time^2 is below −1.6.	For each enumerator, ratio of flagged interviews to completed interviews in that week was calculated. Within each state-region-gender strata, top three enumerators with highest positive ratios (if any) were flagged for intervention.
6	Section 4 Time	Effort	Time taken to field 'Weekly Labour Force Status'.	Very short section time suggests violation of interview protocols.	Paradata	Interview got flagged if its standardized section time^2 is below −1.6.	For each enumerator, ratio of flagged interviews to completed interviews in that week was calculated. Within each state-region-gender strata, top three enumerators with highest positive ratios (if any) were flagged for intervention.

TABLE 3 (Continued)

S. no.	Flag name	Dimension being evaluated	Field practice being monitored	Underlying concern	Based on paradata or survey data	Threshold	Ranking enumerators for intervention
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
7	Section 5 Time	Effort	Time taken to field 'Household Production Activities'.	Very short section time suggests violation of interview protocols.	Paradata	Interview got flagged if its standardized section time^2 is below −1.6.	For each enumerator, ratio of flagged interviews to completed interviews in that week was calculated. Within each state-region-gender strata, top three enumerators with highest positive ratios (if any) were flagged for intervention.
8	Section 8 Time	Effort	Time taken to field 'Decision Making'.	Very short section time suggests violation of interview protocols.	Paradata	Interview got flagged if its standardized section time^2 is below −1.6.	For each enumerator, ratio of flagged interviews to completed interviews in that week was calculated. Within each state-region-gender strata, top three enumerators with highest positive ratios (if any) were flagged for intervention.
9	Section 9 Time	Effort	Time taken to field 'Intergenerational Mobility'.	Very short section time suggests violation of interview protocols.	Paradata	Interview got flagged if its standardized section time^2 is below −1.6.	For each enumerator, ratio of flagged interviews to completed interviews in that week was calculated. Within each state-region-gender strata, top three enumerators with highest positive ratios (if any) were flagged for intervention.

(Continues)

TABLE 3 (Continued)

S. no.	Flag name	Dimension being evaluated	Field practice being monitored	Underlying concern	Based on paradata or survey data	Threshold	Ranking enumerators for intervention
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
10	Section 10 Time	Effort	Time taken to field 'Social Networks'.	Very short section time suggests violation of interview protocols.	Paradata	Interview got flagged if its standardized section time[2] is below −1.6.	For each enumerator, ratio of flagged interviews to completed interviews in that week was calculated. Within each state-region-gender strata, top three enumerators with highest positive ratios (if any) were flagged for intervention.
11	Roster Size	Ethics	Number of household members listed in the household roster.	If enumerator deliberately leaves out some household members, it may adversely affect survey representativeness.	Survey data	Enumerator got flagged when their standardized average roster size is below −1.[3]	Intervene on all flagged enumerators (if any).
12	Network Size	Content	Number of persons listed in respondent's social network.	If enumerator does not capture everyone in the respondent's network, it may bias the network structure.	Survey data	Enumerator got flagged when their standardized average network size is below −1.[3]	Intervene on all flagged enumerators (if any).
13	Odd Start	Ethics	Whether interview started outside usual survey hours.	Indicates interview falsification.	Paradata	Interview got flagged if its start time was before 6 am or after 9 pm.	Intervene on all enumerators (if any) with at least one flagged interview.

TABLE 3 (Continued)

S. no.	Flag name	Dimension being evaluated	Field practice being monitored	Underlying concern	Based on paradata or survey data	Threshold	Ranking enumerators for intervention
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
14	Alone Section 7	Ethics	Whether an enumerator reports that the respondent was interviewed in private when fielding the Discrimination section.	If enumerator reports being always alone or being never alone, they are perhaps not documenting the privacy status correctly.	Survey data	Enumerator got flagged if for all completed interviews in that week they recorded either being always alone or never alone with the respondent.	Intervene on all flagged enumerators (if any).
15	Section 4 Skip	Content, Effort, Ethics	Whether respondent's work status is reported as 'Not Working'.	If enumerator is not clear about what constitutes work, or does not probe enough, or deliberately records 'not working', it may bias estimates of work status.	Survey data	Interview got flagged if respondent was reported as 'not working'.	For each enumerator, ratio of flagged interviews to completed interviews in that week was calculated. Within each state-region-gender strata, top three enumerators with highest positive ratios (if any) were flagged for intervention.

[1]The IWS questionnaire has 13 sections. A section was included in the calculation of 'Survey Time' only if a) it was fielded to ALL respondents, b) AND its anticipated length was NOT linked to the respondent's gender/work profile. The sections that were included are Demographic Characteristics, Household Production, Discrimination, Decision Making, and Networks.

[2]Flags for survey/section times were created at the interview level. Standardized survey/section times at the interview level were calculated as corresponding z-scores created using the mean and standard deviation across all interviews conducted in the enumerator's comparison group in that week.

[3]Flags for 'Roster Size' and 'Network Size' were created at the enumerator level. First, each enumerator's average size was calculated using all interviews completed by them in that week. Next, standardized average sizes at the enumerator level were calculated as corresponding z-scores created using the mean and standard deviation across all enumerators in the enumerator's comparison group, in that week.

flags that are hard to talk about. In retrospect, we recommend using only two time-based flags, namely those capturing the overall interview time and the time spent on the core module of the survey. This is because time-based flags tend to be positively correlated and each additional flag contributes only marginally in terms of incremental information. Supplementary Appendix A provides a rationale for using "Survey Time" in lieu of other time-based flags.

4 | INTERVENING TO IMPROVE SURVEY QUALITY

Next, we describe our interventions on flagged enumerators.

4.1 | Timeline of interventions

Flags were collated into weekly reports, one for each state. Supplementary Appendix S1 presents a sample report. These reports were shared with the respective state-level manager who in turn emailed them to all field supervisors. The manager followed this up with a phone call to those field supervisors who had at least one flagged enumerator and highlighted the issues against each flagged enumerator. The final step involved a private conversation between a field supervisor and a flagged enumerator, where the supervisor would point out the deviant behaviour and nudge the enumerator to take corrective action without being accusatory.

Two reports were shared with the IWS field personnel. The first was based on enumerator performance in the week between February 17 and February 23, and the second between February 24 and March 8.[7] The first report was shared on March 3 and March 4 in Rajasthan and Karnataka, respectively; while the second was shared on March 10 and March 14, in Rajasthan and Karnataka, respectively. We examine interventions based on only the first report and disregard the second one because: (a) around the time of sharing the second report, COVID-19 was beginning to impact enumerator psyche, making it impossible to disentangle the effect of COVID-19 from that of our interventions; and (b) the second report is likely to interact with the first, and one cannot separate the independent effects of the two reports once the second report has been shared.

4.2 | Effectiveness of interventions

We use ordinary least squares regressions with enumerator fixed effects to analyse the impact of interventions based on the first report. We estimate the following equation:

$$Performance_{ij}^{k} = \beta_0 + \beta_1\left(FlagSame_j^k * Post_i\right) + \beta_2\left(FlagOther_j^k * Post_i\right) + \beta_3\left(Date_i\right) + \beta_4\left(DateSquared_i\right) + \left\{Enumerator_i\right\} + \varepsilon_{ij}^{k} \quad (1)$$

i is for interview, j for enumerator, and k for a specific flag such as "Survey Time" or "Section 4 Skip" (column (2) of Table 3 lists all the flags). Recall that each flag is associated with a specific field practice. $Performance^k$ refers to the measurement of the underlying field practice associated with flag k. For example, for the Survey Time flag, the performance measure is the interview duration in minutes and for Section 4 Skip it is a 0/1 indicator for whether

[7]The first two weeks of the survey were not targeted for intervention because processes are typically in flux in the initial period and it takes a while before they stabilize.

or not the respondent was reported as "Not Working."[8] *FlagSame*^{k} and *FlagOther*^{k} are indicators for whether the enumerator was flagged for flag *k* and for some other flag (other than *k*), respectively. *Post* is an indicator for whether the interview was closed in the post-intervention period. *Date* and *DateSquared* form a quadratic in time and {*Enumerator*} is the set of enumerator fixed effects.[9] ε captures all idiosyncratic factors that affect performance. To improve precision, we only included enumerators with at least 10 completed interviews in all our regressions and clustered standard errors at the enumerator level, which is also the level at which intervention happens (Abadie et al., 2022).

The primary coefficient of interest is β_1. It captures the change in performance as a result of field supervisors talking to flagged enumerators. β_2 is also of interest and shows whether intervening to correct some other practice had an effect. The time controls, *Date* and *DateSquared*, account for secular changes that affect all enumerators. Finally, by including enumerator fixed effects we are identifying the effect of interventions by looking at whether an enumerator changed behaviour relative to their own behaviour prior to being flagged. This makes it more likely that β_1 is capturing the causal effect of intervening and is not being influenced by systematic personality differences between flagged and other enumerators.[10]

Table 4 presents descriptive statistics on flags along with regression results. Recall that the first report flagged enumerators based on their performance from February 17 to February 23. The table examines a longer period between February 17 and March 10 for Karnataka and between February 17 and March 14 for Rajasthan. We refer to this as the analysis period. Of this, the pre-intervention period is before March 5 and March 6 for Rajasthan and Karnataka, respectively. During the analysis period, a total of 88 enumerators completed at least one interview, of which 46 were with women. The only flags studied are those for which at least one enumerator was flagged. Column (3) shows the number of enumerators flagged against each flag. Columns (4) and (5) present the mean value of the performance measure during the pre-intervention period for "all" and "flagged" enumerators, respectively. The regression results are shown in columns (6) through (10). Columns (6) and (7) present our estimates for β_1 and β_2, respectively.

A look at our main coefficient, β_1, shows that our interventions had the intended effect for one crucial flag, namely "Section 4 Time": the interview time for Section 4 increased by 0.7 minutes or 18% of the pre-intervention average time for this section. At the same time, the interview time for section 8 reduced by 0.2 minutes, see "Section 8 Time." Given that Section 4 constitutes the core of IWS, it is very likely that field supervisors emphasized it the most during feedback sessions. It is therefore not surprising if flagged enumerators focused on Section 4 at the cost of other sections. We conjecture that the observed effects are explained by (a) enumerators compensating for increased time in Section 4 by cutting back on time spent elsewhere to meet the agency's productivity targets; and (b) too many flags adversely affecting communication between supervisors and flagged enumerators. Our second conjecture is partly strengthened by some significant estimates for β_2: flagging for some other practice increased the time spent on Sections 4 (Section 4 Time) and 5 (Section 5 Time) and lowered the number of respondents reported as "Not Working" (Section 4 Skip), though the last result is statistically significant only at the 10% level.

In light of the finding that our interventions based on Section 4 Time resulted in increased time spent fielding that section, we examine whether this increased time translated into substantive improvements in data quality. We do this by looking at enumerators who were flagged for either Survey Time or Section 4 Time and comparing

[8]For all except three flags, *Performance*^{k} is a continuous variable. For Odd Start, Alone Section7, and Section4 Skip it is binary. We estimated a probit specification for the three binary variables and found the substantive results did not change.

[9]In the presence of a quadratic time trend and enumerator fixed effects, we cannot include *FlagSame* and *FlagOther* as standalone explanatory variables due to perfect multicollinearity.

[10]An alternative to enumerator fixed effects is to treat them as random effects uncorrelated with ϵ. The Hausman test rejects the random effects model for three flags, namely Survey Time, Network Size, and Alone Section7. Moreover, in case of Survey Time, the coefficient on *FlagSame* flips sign between the fixed and random effects models. We prefer the fixed effects specification as it is consistent under both the null and alternative hypotheses of the Hausman test.

TABLE 4 Effect of Paradata-Based Interventions on Enumerator Performance.

		Descriptive Statistics			Regression Results				
S. No.	Flag Name	Number of Flagged Enumerators of Enumerators with at least 1 completed interview	Mean over Interviews of All Enumerators in Pre-Intervention Period	Mean over Interviews of Flagged Enumerators in Pre-Intervention Period	Coefficient value, Enumerator Flagged for Same Field Practice	Coefficient value, Enumerator Flagged for at least one Other Field Practice	Number of Flagged Enumerators of Enumerators with at least 10 completed interviews	R squared	Number of Observations/ Completed Interviews (Number of Clusters/ Enumerators)
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
1	Survey Time (minutes)	12 of 88	14.5	12.2	0.029	0.614	12 of 75	0.27	3013 (75)
			(5.8)	(4.3)	(0.820)	(0.466)			
2	Section 2 Time (minutes)	5 of 46[1]	2.2	2.1	0.463	0.104	5 of 39[1]	0.18	1779 (39)
			(0.9)	(1.0)	(0.517)	(0.112)			
3	Section 4 Time (minutes)	2 of 88	4.1	2.3	0.732***	0.370**	1 of 75	0.05	3009 (75)
			(6.4)	(1.8)	(0.165)	(0.184)			
4	Section 5 Time (minutes)	1 of 88	1.9	1.6	−0.025	0.228**	1 of 75	0.21	3009 (75)
			(1.1)	(1.1)	(0.088)	(0.096)			
5	Section 8 Time (minutes)	1 of 88	1.6	2.1[2]	−0.237***	0.047	1 of 75	0.18	3009 (75)
			(1.2)	(0.8)	(0.071)	(0.102)			
6	Section 9 Time (minutes)	6 of 88	2.0	1.8	0.134	−0.092	5 of 75	0.15	3008 (75)
			(1.0)	(0.9)	(0.187)	(0.077)			
7	Section 10 Time (minutes)	7 of 88	4.2	3.8	0.059	0.048	5 of 75	0.24	3009 (75)
			(2.5)	(2.3)	(0.262)	(0.254)			

TABLE 4 (Continued)

		Descriptive Statistics			Regression Results				
S. No.	Flag Name	Number of Flagged Enumerators of Enumerators with at least 1 completed interview	Mean over Interviews of All Enumerators in Pre-Intervention Period	Mean over Interviews of Flagged Enumerators in Pre-Intervention Period	Coefficient value, Enumerator Flagged for Same Field Practice	Coefficient value, Enumerator Flagged for at least one Other Field Practice	Number of Flagged Enumerators of Enumerators with at least 10 completed interviews	R squared	Number of Observations/ Completed Interviews (Number of Clusters/ Enumerators)
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
8	Network Size (members)	7 of 88	3.2	2.4	0.201	−0.006	6 of 75	0.47	2994 (75)
			(1.5)	(1.2)	(0.474)	(0.112)			
9	Alone Section 7 (1 if alone, 0 otherwise)	37 of 88	0.87	0.93	−0.052	0.013	36 of 75	0.35	3009 (75)
					(0.068)	(0.060)			
10	Section 4 Skip (1 if follow up section not needed, 0 otherwise)	12 of 88	0.32	0.42	−0.071	−0.070*	11 of 75	0.17	3009 (75)
					(0.044)	(0.041)			

The table examines the set of interventions based on the first report which covered enumerator performance in the week from February 17 to February 23. Section 0 Time, Section 1 Time, Section 3 Time, Roster Size, and Odd Start are omitted from the table, as none of the enumerators were flagged for these in the first report. The descriptive statistics are limited to enumerators who completed at least one interview (following the strict definition of a completed interview), while the regressions are limited to enumerators who completed at least 10 interviews. The regression analysis is based on enumerator performance between February 17 and March 10 for Karnataka, and between February 17 and March 14 for Rajasthan. Regressions are at the interview level and the dependent variable is indicated under the column (2) Flag Name. For descriptive statistics in columns [4] and [5], standard deviations are shown in parentheses. For regression coefficients in columns [6] and [7], clustered standard errors, clustered at the enumerator level, are shown in parentheses. * stands for statistical significance at the 10% level of significance, ** at 5%, and *** at 1%.

[1]Section 2 was administered only by female enumerators.

[2]The mean for flagged enumerators could be higher than the mean for all enumerators because flags were generated based on performance between February 17–23, whereas the means shown in the table are based on performance over a longer pre-intervention period which starts on February 17 and goes all the way to the date of intervention (March 4 for Rajasthan and March 5 in Karnataka).

TABLE 5 Select Data Quality Indicators for Flagged Enumerators Before and After Intervention.

S.No.		Pre-Intervention	Post-Intervention
1	Share in Paid Self-employment (%)	27.21	34.00
2	Variance of 2-digit Industry Code given Paid Self-employment	24.13	27.19
3	Variance of 2-digit occupation code given Paid Self-employment	4.85	4.96
4	Share in Unpaid Self-employment (%)	37.19	37.00
5	Variance of 2-digit Industry Code given Unpaid Self-employment	11.87	13.13
6	Variance of 2-digit occupation code given Unpaid Self-employment	1.32	2.34
7	Share in Paid Wage Work (%)	19.73	18.00
8	Variance of 2-digit Industry Code given Paid Wage Work	32.56	35.32
9	Variance of 2-digit Occupation Code given Paid Wage Work	5.75	6.84
10	Number of Interviews	441	100

The table is based on performance measures of all flagged enumerators (flagged for Survey Time or Section 4 Time) who interviewed in both the pre- and post-intervention period. There are 12 such enumerators in all.

their average performance in terms of select data quality measures from Section 4, before and after intervening.[11] The results are shown in Table 5. Higher shares of respondents reported that they were working, either in self-employment (paid or unpaid) or in wage work, automatically imply that additional details about their work would need to be collected. This requires more effort on the part of the enumerator.

We see in Table 5 that our interventions are associated with a large increase in the share of those reported as working in paid self-employment (about 7 percentage points); accompanied by some decline in shares reported as working in unpaid self-employment and paid wage work, these declines are much smaller (less than 1 and 2 percentage points, respectively). A higher variance of industry and occupational codes in the post-intervention period also suggests that the enumerators expended more effort to correctly document these aspects rather than using the same set of codes for everyone.

In summary, Table 5 suggests that our interventions improved data quality. In Supplementary Appendix C, we look for heterogeneity in enumerator response to interventions based on their pre-survey characteristics. None of the enumerator characteristics we examine, namely, age, gender, education or performance in a pre-survey test, can explain the improvements seen in Table 5.

We have shown that our interventions made a difference and changed enumerator behaviour in the field. For one crucial practice, namely the time spent fielding section 4, the change was along intended lines, while for a few others it was not. We believe that better training for field supervisors with a focus on how to effectively communicate with flagged enumerators and reducing the number of flags should mitigate some of the unintended effects.

[11]We did not limit this exercise to only those enumerators who were flagged for Section 4 because, as Table 4 shows, there were only two such enumerators. Instead, we look at all enumerators who were either flagged for shorter overall survey time (Survey Time Flag) or for shorter Section 4 time (Section4 Flag). This is sensible because, as mentioned earlier, Section 4 is the main module of the survey which makes it highly likely that it was emphasized during conversations with enumerators flagged for both these performance measures.

5 | LESSONS LEARNT FROM IWS

5.1 | Monitoring overall progress of the survey

In any survey two objectives compete: completing a fixed number of interviews per day to avoid cost over-runs; versus enumerators spending adequate time with each respondent to ensure meaningful data are being collected. To strike a balance between these competing demands, we recommend tracking three parameters at the survey level based on paradata, namely: the count of completed interviews; average time per completed interview; and the ratio of completed to initiated interviews. These parameters are necessary and sufficient to strike a balance between monitoring many important aspects of survey implementation and tracking too many measures that result in obfuscation of information.

Next, we discuss each parameter, and Supplementary Appendix D shows their evolution over the course of IWS.

5.1.1 | Count of completed interviews

A key issue often overlooked by PIs is defining what constitutes a completed interview. From an agency's perspective, an interview is complete when the enumerator has gone over all the relevant sections with the respondent; whereas for PIs the nature of non-response also matters. For this reason, we recommend a stricter definition: an interview is complete only when the enumerator has spent a minimum amount of time on each mandatory section administered to all respondents. We recommend that PIs use this stricter definition to track completed interviews, as it would typically give a more conservative picture of how data collection is progressing.

5.1.2 | Average time per completed interview

Tracking the average time per completed interview is established best practice. If the average time falls below a critical value, it indicates that less time is being spent with respondents which may compromise the quality of data.

5.1.3 | Ratio of completed to initiated interviews

A low value ratio of completed to initiated interviews suggests futility of effort by enumerators. Once initiated, an interview could end up being incomplete for multiple reasons—such as the respondent not giving consent or withdrawing it later on; respondents stopping the survey midway through; respondents not being available during revisits; or the interview not meeting the minimum time described in subsection 5.1.1. Improving this ratio may entail one or more of the following: instituting a more effective consent delivery; checking with the respondent before fixing revisit times; rethinking the revisit protocol and stopping rule (maximum number of revisits before an interview is closed); and sensitizing enumerators to spend more time with each respondent.

5.2 | Understanding paradata composition beforehand

To avoid scrutiny of its operations, the data collection agency may not always be forthcoming in sharing detailed paradata. It is important to know the exact variables that the agency is willing and able to share with the PIs. To set feasible flags, one must know the granularity of time stamps, whether they are at the interview, section, or question level. Ideally, paradata requirements should be included in the contract between PIs and the agency.

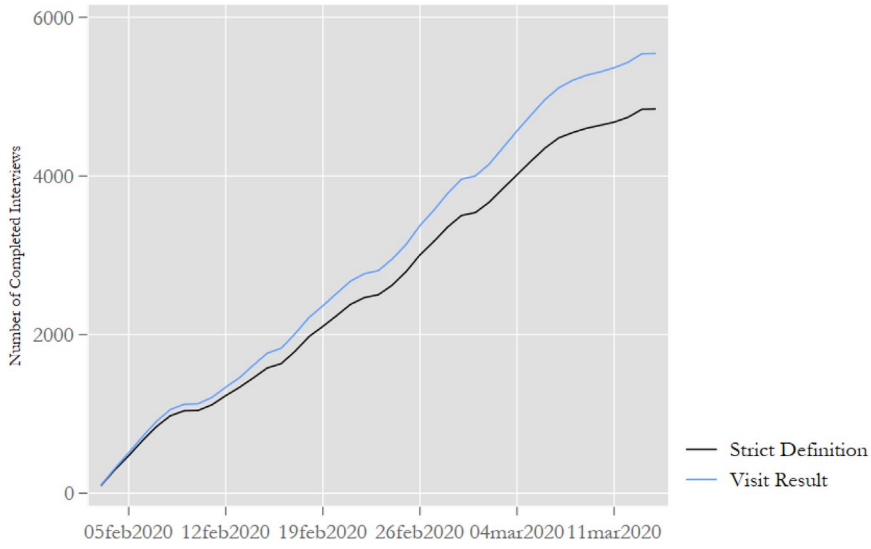

FIGURE 1 Cumulative Count of Completed Interviews.

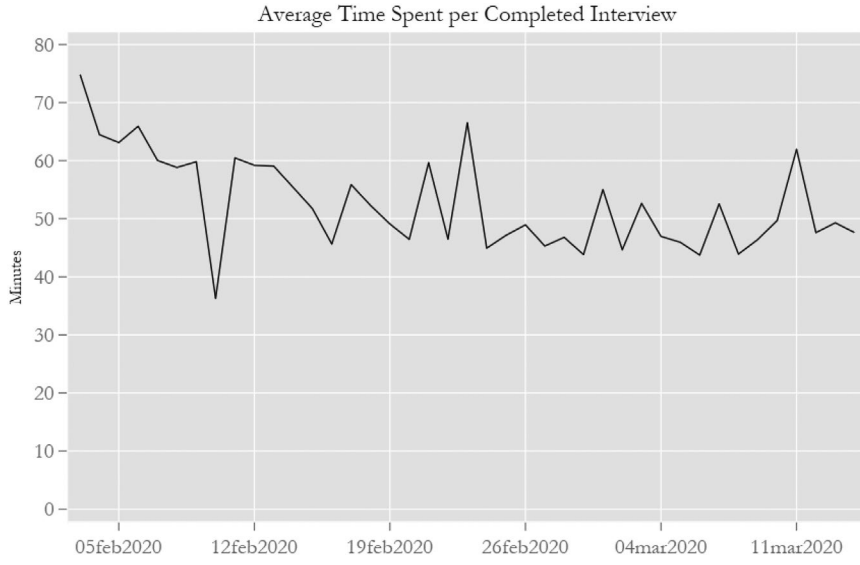

FIGURE 2 Figure 2 shows the average time per completed interview over the course of IWS using the stricter definition of completed interviews. Typically, as enumerators become increasingly adept at administering the survey the average interview time first drops and then stabilises. In IWS it took about two weeks for the interview time to stabilise: the average interview time was 61 minutes in the first two weeks and reduced to 48 minutes thereafter. The variance of interview times was large: 44 minutes in the first two weeks and 37 minutes thereafter.

5.3 | Dashboard versus manually generated reports

A dashboard is a one-stop-shop where key performance indicators can be seen at a glance. Recent literature recommends using dashboards for performance management (Sarikaya et al., 2019). While dashboards that generate

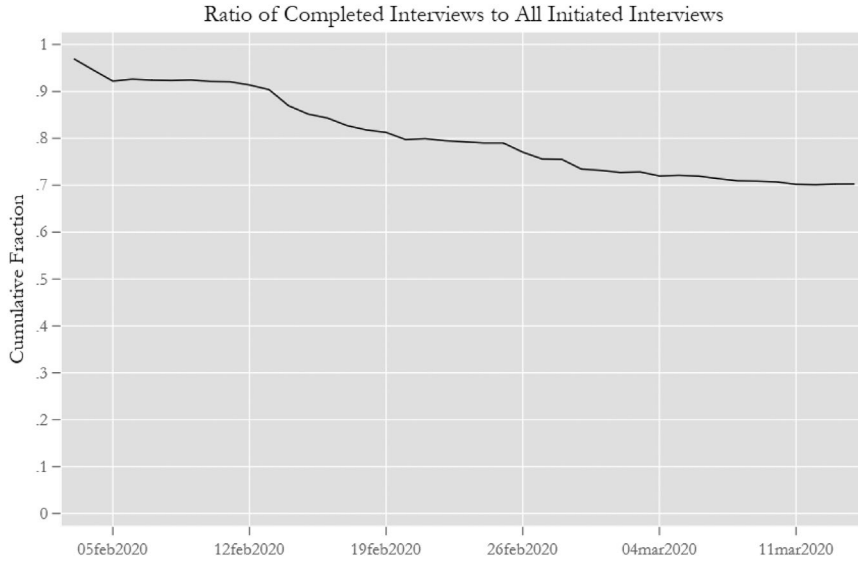

FIGURE 3 Figure 3 shows the ratio of completed to initiated interviews over the course of IWS, again using the stricter definition of completed interviews. The decline in this ratio over the course of the survey is largely due to an extraneous factor beyond our control, namely the nation-wide protests against the Citizenship Amendment Act (CAA) that were gaining momentum at that time.[13]

automated reports in real time are undoubtedly preferable over a system that generates manual reports with a lag, they may not always be feasible when resources are limited.

On the one hand, using freeware to design a dashboard from scratch requires specialized coding skills which may be hard to find; on the other hand, using paid applications that come with in-built customization may be too expensive. When resources are scarce, manual reports may be better than a poorly designed dashboard. In IWS we used a dashboard to track overall progress of the survey,[12] but relied on manually generated reports for sharing information about flagged enumerators. Supplementary Appendix E shows screenshots of the dashboard we used.

6 | CONCLUDING REMARKS

We describe how we used paradata to improve the quality of the IWS survey while the survey was still in progress. Causal regression analysis shows our method of intervening using paradata-based flags tempered deviant enumerator behaviour.

A crucial aspect of our method is dynamic benchmarking among enumerators in the same comparison group consisting of enumerators who faced similar field conditions. Because of this, whenever our method reduces variance in field practices, it should translate into lower standard errors for estimates based on subsamples demarcated by dimensions used to define comparison groups. In IWS, the three delineating dimensions were state, subregion, and gender. Any IWS estimate specific to a state, subregion, or gender (or to any

[12]We designed the IWS dashboard using Shiny, an open-source R-based package for building web applications.

[13]The CAA allows for non-Indian individuals from certain religious communities, in select countries, to become Indian citizens. Controversially, it excludes Muslims from the eligible list. Given that IWS focuses on religious identity, respondents, especially Muslims, were increasingly fearful of participating in the survey, resulting in declining completion rates over time.

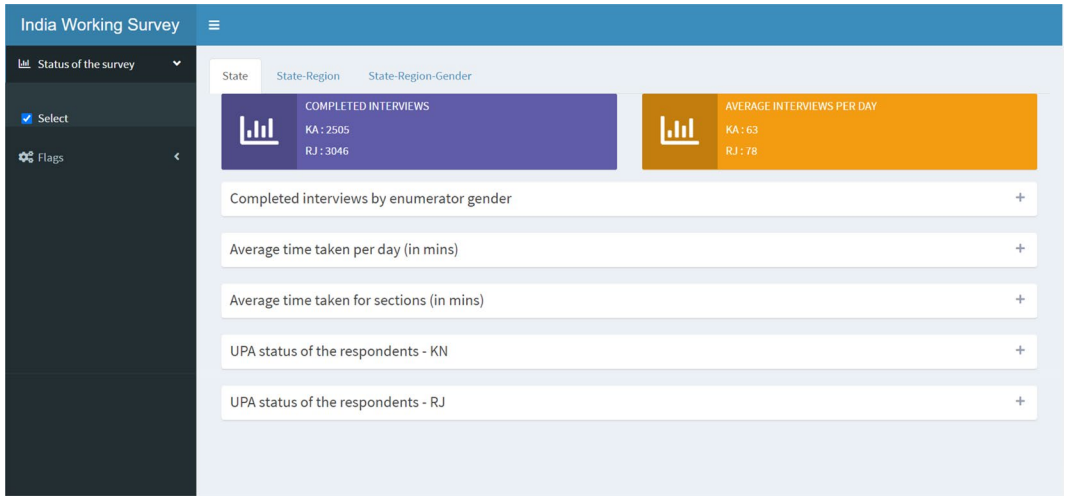

FIGURE 4 Outer layer.

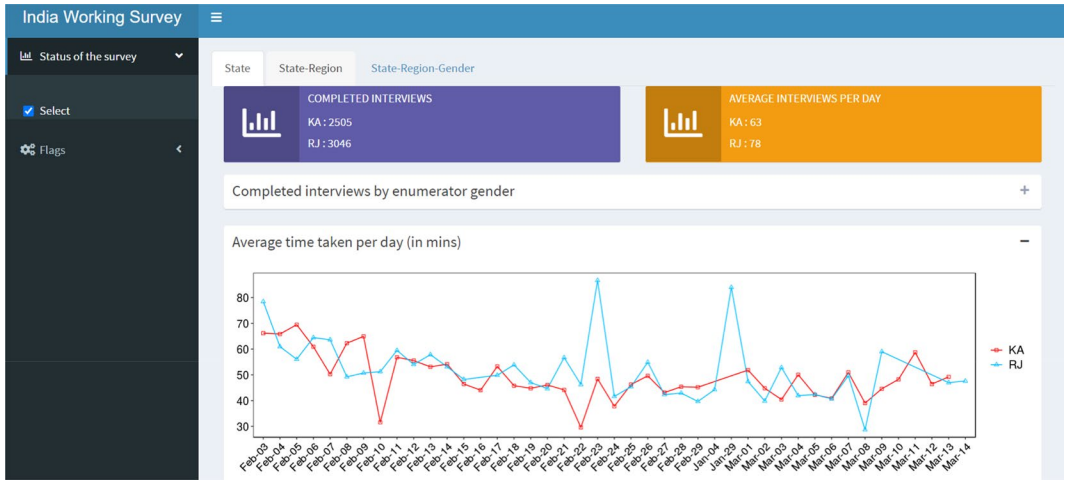

FIGURE 5 Inner layer.

combination of these) should have a lower standard error compared to the counterfactual where such quality control was absent.

When our method succeeds, data quality improves during the survey owing to greater coherence in field practices over time. What then should be done with data collected early in the survey? This is best left to the discretion of individual survey users. Ideally, all deviant observations should be identified in the dataset. Empirical evidence will be considered reliable only if the main results are immune to whether or not deviant observations are included in the estimation, bolstering the credibility revolution within economics (Angrist & Pischke, 2010).

Although paradata have tremendous potential to improve survey quality, they remain underused especially in LMICs. One way to encourage their use is for donor agencies funding surveys to: a) mandate paradata use; b) earmark a part of the overall budget for paradata-based quality control; and c) require that some paradata (e.g. interview length) be made public along with the survey data. This could make paradata use a standard practice worldwide. For-profit data collection agencies will then view paradata not as a threat to their commercial interests but as an integral tool to improve their business Figures 1–5.

AUTHOR CONTRIBUTIONS

Both authors worked jointly and in equal measure towards building the conceptual underpinnings, framework or research design/methodology of the given publication; data collection and processing; analysis and interpretation of the data. The corresponding author played a lead role in writing the paper and substantively and substantially revising it for rigour and coherence.

The authors have no competing interests to declare.

ACKNOWLEDGEMENTS

We are grateful to the National Council of Applied Economic Research (NCAER), New Delhi, for funding this research. We are grateful to Santanu Pramanik and Sonalde Desai for their comments and feedback. We thank the Institute for What Works to Advance Gender Equality, Azim Premji University, and the Indian Institute of Management Bangalore for funding the India Working Survey on which this article is based. All the views presented here are those of the authors and not of any of the institutes mentioned above. We thank Rahul Lahoti for help in securing the funds for this survey, involvement in the collection of paradata and survey data, and for comments on earlier drafts. The authors acknowledge research assistance from Naveen Gajjalagari and Mridhula Mohan. All errors and omissions are their own.

DATA AVAILABILITY STATEMENT

The datasets generated during and/or analysed during the current study are not publicly available. However, they are available from the corresponding author on reasonable request.

ORCID


Rosa Abraham https://orcid.org/0000-0003-3306-9621

REFERENCES

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2022). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, *138*(1), 1–35. https://doi.org/10.1093/qje/qjac038

Ackermann-Piek, D., Korbmacher, J. M., & Krieger, U. (2020). Explaining interviewer effects on survey unit nonresponse: A cross-survey analysis. In K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, & B. T. West (Eds.), *Interviewer effects from a total survey error perspective* (1st ed., pp. 193–206). Chapman and Hall/CRC. https://doi.org/10.1201/9781003020219-19

Angrist, J. D., & Pischke, J. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *The Journal of Economic Perspectives*, *24*(2), 3–30. https://doi.org/10.1257/jep.24.2.3

Bhuiyan, M. F., & Lackie, P. (2017). Mitigating survey fraud and human error: Lessons learned from a low budget village census in Bangladesh. *IASSIST Quarterly*, *40*(3), 20–26. https://iassistquarterly.com/public/pdfs/vol_40-3_20_26.pdf

Bushery, J. M., Reichert, J. W., Albright, K. A., & Rossiter, J. C. (1999). Using date and time stamps to detect interviewer falsification. *Proceedings of the survey research method section*, Data Quality in Surveys Section (pp. 316–320). American Statistical Association. http://www.asasrms.org/Proceedings/y1999f.html

Choumert-Nkolo, J., Cust, H., & Taylor, C. (2019). Using paradata to collect better survey data: Evidence from a household survey in Tanzania. *Review of Development Economics*, *23*(2), 598–618. https://doi.org/10.1111/rode.12583

Couper, M. (1998). Measuring survey quality in a CASIC environment. In *Proceedings of the Survey Research Methods Section of the ASA at JSM 1998*, Achieving Quality in Surveys Section (pp. 41–49). American Statistical Association. http://www.asasrms.org/Proceedings/y1998f.html

Da Silva, D. N., & Skinner, C. J. (2020). Testing for measurement error in survey data analysis using paradata. *Biometrika*, *108*(1), 239–246. https://doi.org/10.1093/biomet/asaa050

Edwards, B., Maitland, A., & Connor, S. (2017). Measurement error in survey operations management. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, & B. T. West (Eds.), *Total survey error in practice* (pp. 253–277). John Wiley & Sons. https://doi.org/10.1002/9781119041702.ch12

Edwards, B., Sun, H., & Hubbard, R. (2020). Behavior change techniques for reducing interviewer contributions to total survey error. In K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, and B. T. West (Eds.), *Interviewer effects*

from a total survey error perspective (1st ed pp. 77–89). Chapman and Hall/CRC. https://doi.org/10.1201/9781003020219-9

Finn, A., & Ranchhod, V. (2017). Genuine fakes: The prevalence and implications of data fabrication in a large South African survey. *The World Bank Economic Review*, *31*(1), 129–157. https://doi.org/10.1093/wber/lhv054.

Guyer, H. M., West, B. T., & Chang, W. (2021). *The interviewer performance profile (IPP): A paradata driven tool for monitoring and managing interviewer performance*. Survey Methods: Insights from the Field. https://doi.org/10.13094/SMIF-2021-00005

Hood, C. C., & Bushery, J. M. (1997). Getting more bang from the reinterview buck: Identifying "at risk" interviewers. *Proceedings of the survey research method section*, Validating the Accuracy of Survey Data Section (pp. 820–824). American Statistics Association. http://www.asasrms.org/Proceedings/y1997f.html

Jans, M., Sirkis, S., & Morgan, D. (2013). Managing data quality indicators with paradata based statistical quality control tools: The keys to survey performance. In F. Krueter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 191–229). Wiley. https://doi.org/10.1002/9781118596869.ch9

Kosyakova, Y., Olbrich, L., Sakshaug, J., & Schwanhäuser, S. (2019). *Identification of interviewer falsification in the IAB-BAMF-SOEP survey of refugees in Germany* (FDZ Method Report No. 02/2019). https://doi.org/10.5164/IAB.FDZM.1902.en.v1

Kreuter, F., Couper, M., & Lyberg, L. (2010). The use of paradata to monitor and manage survey data collection. In *Proceedings of the joint statistical meetings, Evaluation, Modeling, and Management of Costs and Risks During Changes in Survey Procedures Section* (pp. 282–296). American Statistical Association. http://www.asasrms.org/Proceedings/y2010/Files/306107_55863.pdf

Krueger, B. S., & West, B. T. (2014). Assessing the potential of paradata and other auxiliary data for nonresponse adjustments. *Public Opinion Quarterly*, *78*(4), 795–831. https://doi.org/10.1093/poq/nfu040

Kunz, T., Daikeler, J., & Ackermann-Piek, D. (2023). Interviewer-observed paradata in mixed-mode and innovative data collection. *International Journal of Market Research*, *66*(1), 14–26. https://doi.org/10.1177/14707853231184742

Kuriakose, N., & Robbins, M. (2018). Don't get duped: Fraud through duplication in public opinion surveys. *Statistical Journal of the IAOS*, *32*(3), 283–291. https://doi.org/10.3233/sji-160978

Lupu, N., & Michelitch, K. (2018). Advances in survey methods for the developing world. *Annual Review of Political Science*, *21*(1), 195–214. https://doi.org/10.1146/annurev-polisci-052115-021432

Mohadjer, L., & Edwards, B. (2018). Paradata and dashboards in PIAAC. *Quality Assurance in Education*, *26*(2), 263–277. https://doi.org/10.1108/qae-06-2017-0031

Olson, K., Smyth, J. D., Dykema, J., Holbrook, A. L., Kreuter, F., & West, B. T. (Eds.) (2020). Interviewer effects from a total survey error perspective. *Chapman and Hall/CRC*. https://doi.org/10.1201/9781003020219

Pashazadeh, F., Cernat, A., & Sakshaug, J. W. (2020). Investigating the use of nurse paradata in understanding nonresponse to biological data collection. In K. Olson, J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, & B. T. West (Eds.), *Interviewer effects from a total survey error perspective* (1st ed., pp. 221–234). Chapman and Hall/CRC. https://doi.org/10.1201/9781003020219-21

Sarikaya, A., Correll, M., Bartram, L., Tory, M., & Fisher, D. (2018). What do we talk about when we talk about dashboards? *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 682–692. https://doi.org/10.1109/tvcg.2018.2864903

West, B. T., & Groves, R. M. (2013). A propensity-adjusted interviewer performance indicator. *Public Opinion Quarterly*, *77*(1), 352–374. https://doi.org/10.1093/poq/nft002

Yan, T., & K. Olson (2013). Analyzing paradata to investigate measurement error. In F. Krueter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 73–95). John Wiley and Sons. https://doi.org/10.1002/9781118596869.ch4

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Goel, D. & Abraham, R. (2025). Improving survey quality using paradata: Lessons from a field survey in India. *Development Policy Review*, *43*, e12813. https://doi.org/10.1111/dpr.12813