I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15041

Improving Survey Quality Using Paradata: Lessons from the India Working Survey

Deepti Goel
Rosa Abraham
Rahul Lahoti

JANUARY 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15041

Improving Survey Quality Using Paradata: Lessons from the India Working Survey

Deepti Goel
Azim Premji University and IZA

Rosa Abraham
Centre for Sustainable Employment at Azim Premji University

Rahul Lahoti
ETH Zurich

JANUARY 2022

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Improving Survey Quality Using Paradata: Lessons from the India Working Survey*

We describe the design and implemention of a paradata based method to reduce interviewer induced measurement error in a household survey in India. Our method identifies enumerators exhibiting deviant field practices, and provides them feedback to correct potentially faulty behavior. A novel feature is the emphasis on dynamic benchmarking within a group of enumerators facing similar field conditions. This helps to correctly pin down steady state levels of multiple data generating processes that exist within our survey. We also present evidence that our method succeeded in changing actual enumerator behavior in the field. Furthermore, we provide a complete prototype of how to operationalize paradata use in a resource constrained environment. At each step, we highlight the trade-offs involved, share insights from our own shortcomings, and provide recommendations to help make more informed choices. We hope our work will encourage the use of paradata to improve survey quality, especially in low- and middle-income countries where their use is still rare.

JEL Classification: C83

Keywords: paradata, survey data, interviewer effects, India

Corresponding author:
Deepti Goel
School of Arts and Sciences
Azim Premji University
Survey No 66, Burugunte Village
Bikkanahalli Main Road
Sarjapura, Bengaluru 562125
India
E-mail: deepti.goel@apu.edu.in

* We are grateful to the National Council of Applied Economic Research (NCAER), New Delhi, for funding this research. The views presented are those of the authors, and not of NCAER, or its Governing Body. We are also grateful to the Initiative for WhatWorks to AdvanceWomen and Girls in the Economy, Azim Premji University, and the Indian Institute of Management (Bangalore), for funding 'The India Working Survey', on which this paper is based on.

1 Introduction

The efficacy of survey-based policy recommendations primarily hinges on the quality of data collected. Does the survey represent the population it claims to characterize? Are respondents voicing their true opinions? Did enumerator bias creep into the data? These are questions that most survey users have, but, are typically brushed aside in the race to get the analyses out. While there are no foolproof measures to guarantee the authenticity of survey data, steps can be taken to improve their credibility. One such, is the use of paradata to streamline enumerator practices. Paradata refers to data about the *process* of data collection (Couper, 1998). Here, we share our experience regarding paradata use to improve the India Working Survey (IWS), a field-based household survey implemented in two Indian states, in 2020.[1]

Paradata usually includes data on who conducted the interview, start- and end-time stamps for the full interview and for individual items, and re-visit information. It could also include keystrokes in case of computer aided interviewing (CAI), global positioning system (GPS) co-ordinates of enumerator movement and interview location, interviewer characteristics, interviewer observations, and audio/video recordings of respondent-interviewer interaction. Paradata collection has been greatly facilitated by CAI, wherein an electronic device prompts the next question based on answers to previous ones. Along with programming the sequence of questions, the device can also be configured to record paradata, making such auxiliary data readily available. The first wave of IWS was conducted as face-to-face interviews using CAI. Here we describe how paradata was used from this wave to, (a) monitor survey progress, and (b) streamline enumerator practices in the field. The ultimate goal was to improve IWS data quality by reducing interviewer induced measurement error, a particularly important component within the Total Survey Error framework (Olson et al., 2020; Schaeffer et al., 2010; West and Blom, 2016).

A rich body of work exists on *post-survey* use of paradata to assess and correct for

[1]IWS has seven principal investigators (PIs), including all three authors of this paper. 'We', variously refers to either all the PIs, or only the three authors.

non-response error (Ackermann-Piek et al., 2020; Krueger and West, 2014; Krueter and Olson, 2013; Pashazadeh et al., 2020), and measurement error (Da Silva and Skinner, 2020; Yan and Olson, 2013). Contrasting this, work on paradata use *concurrent* with survey implementation is still emerging. Edwards et al. (2017, 2020); Mohadjer and Edwards (2018) used anomalies detected in the data, as they were being collected, to provide immediate feedback to interviewers and thereby improve adherence to survey protocols. Our exercise is similar, but, the method to detect anomalies is different (explained in section 5).

Our method is embedded within the *statistical process control* perspective which advocates adopting procedures used in quality control of industrial products for improving ongoing surveys (Kreuter et al., 2010). Hood and Bushery, 1997 and Bushery et al., 1999, are early papers within this perspective. They identified outlier interviews in the U.S. Census Bureau's household surveys for focused re-interviewing. In more recent work, Kosyakova et al. (2019) used statistical techniques to successfully identify a confirmed fraudulent interviewer in actual survey data from Germany. Although within the statistical process control paradigm, these papers test the effectiveness of different fraud detection methods after the survey is complete, while our method aims to improve data quality of an on-going survey. Moreover, the way they benchmark processes to identify deviant behavior is very different from what we do.[2] Guyer et al. (2021) extend the work by West and Groves (2013) to develop a paradata driven tool for managing interviewer performance in real time. Their tool is very close in spirit to what we discuss here. However, a crucial methodological aspect that we emphasize, namely, dynamic benchmarking within a group of enumerators who face similar external environments, receives only a passing mention in their paper. The conceptual underpinnings of our method can be found in Jans et al., 2013, where they are careful to distinguish special from common cause variation, a key feature of our method as well.

In the last decade or so, household surveys managed by small teams of individual re-

[2]For benchmarking, Hood and Bushery (1997) used covariates from a previous Census, Bushery et al. (1999) used historical survey averages, and Kosyakova et al. (2019) used averages across all interviews. As explained in section 5, we use paradata and survey data as they are being collected to create benchmarks that evolve over time within homogenous groups of enumerators.

searchers have gained traction in many developing countries, including India (Lupu and Michelitch, 2018). Given the sheer volume of critical tasks to be completed before the start of any survey, designing effective use of paradata tends to take a back seat, especially when compounded with severe budget, time, and skilled manpower constraints, a scenario more common in developing countries. This is particularly worrisome as there is suggestive evidence of higher survey fraud in these countries (Kuriakose and Robbins, 2018). It is against this backdrop that our contribution is particularly important. Barring some recent work (Bhuiyan and Lackie, 2016; Choumert-Nkolo et al., 2019; Finn and Ranchhod, 2015), most illustrations of paradata use (including the ones cited earlier), are from developed countries. Even though the underlying statistical theory is portable across contexts, the operationalization issues in less developed countries are very different (Lupu and Michelitch, 2018). By providing a complete prototype of how to operationalize paradata use in a relatively resource constrained environment, we fill an important gap in the paradata literature. In what follows, we are candid about our shortcomings and oversights, and share the lessons we learnt along the way. In doing so, we hope that others will be encouraged to implement and improve our method.

2 IWS Field Operations

IWS was conducted in the states of 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 planned from February through April, 2020. However, field operations had to be stopped in mid-March due to COVID-19 and the subsequent national lockdown. In this paper, we analyze paradata and survey data from this first wave. Appendix A presents the section-wise organization of the field questionnaire, useful for understanding subsequent analyses.

Data collection was outsourced to a private agency. The agency's personnel comprised of 3 senior managers, 15 field supervisors, and around 100 enumerators. The agency's supervisors and enumerators administered the survey by means of computer-aided per-

sonal interviews (CAPI). Additionally, the principal investigators (PIs) directly employed 2 project managers, and 4 independent supervisors, to oversee operations and liaison with the agency.

A total of 6,900 respondents from 3,623 households were contacted between February 3 and March 17, 2020. Every household was visited by one female and one male enumerator, sometimes accompanied by their field supervisor. In keeping with local norms, and given the gender-sensitive nature of some questions, female respondents were only interviewed by female enumerators, and likewise for males.

3 IWS Paradata

Throughout the survey period, we would receive data dumps from the agency every two or three days. Each dump consisted of two files, which contained the paradata and survey data up to that point. In both files, there is a one-to-one correspondence between an observation/row and a respondent. Even when the interview with a respondent unfolded over multiple sittings, the related respondent data appears as a single observation. We, therefore, use 'respondent observation' and 'interview', interchangeably. Table 1 is an exhaustive list of paradata variables used in this paper, captured at the time of administering the questionnaire.

We hired one programmer to exclusively work on paradata and generate reports based on each data dump. These would be shared with the PIs within a day or two of receiving the dump. Thus, in IWS, paradata based survey monitoring had an in-built delay of about four days. Irrespective of this lag time, we do not advise exclusive reliance on paradata for real-time monitoring: additionally, we used 'WhatsApp', an internet-based application for instant messaging over mobile phones. Moreover, there is no substitute for actual PI presence in the field, at least in the early days of the survey.

4 Paradata Monitoring of Survey Progress

In any survey, a silent tug of war ensues between two objectives: completing a fixed number of interviews per day to avoid cost over-runs, versus, requiring enumerators to spend adequate time with each respondent to ensure that meaningful data is collected. Effective use of paradata can strike a balance between these competing demands. We recommend tracking three survey-level parameters, created using paradata.[3]

In deciding the parameters to track, there is a trade-off between wanting to monitor many different aspects of performance, and tracking too many details resulting in obfuscation of information. We advise PIs to be judicious in their choice and recommend tracking the three parameters described in this section. We consider these necessary and sufficient for monitoring most ongoing surveys.

4.1 Cumulative count of completed interviews

A key issue is the definition of a 'completed' interview itself. From an agency's perspective, an interview is complete as long as the enumerator went over all the relevant sections with the respondent, whereas, for the PIs the nature of non-response within each section also matters. As a case in point, in IWS, the enumerators marked 80 percent of the 6,900 initiated interviews as completed, whereas, a stricter definition that mandated a minimum time for select sections in order to ensure that adequate time was spent with the respondent, resulted in a lower completion rate of 70 percent. Figure 1 shows the cumulative count of completed interviews over the survey period according to both definitions: 'Visit Result' refers to what enumerators marked as completed interviews, and 'Strict Definition' to the one set by PIs that required a minimum duration for select sections. The discrepancy between them reinforces the need to track the right metric. The stricter definition can be easily coded using paradata on section times and we recommend that PIs use it to track completed interviews.

[3]Each state in IWS had its own independent collection team, and so the parameters were created and monitored separately for each state.

4.2 Average time per completed interview

When external agencies implement data collection, there is a greater likelihood that enumerators take shortcuts and compromise on interview protocols in order to meet internal productivity targets. This is because, while the financial burden of not completing the survey on time, wholly or partly, falls on the agency, collecting bad data does not have direct implications for them. We, therefore, recommend close tracking of the average interview time throughout the survey period. Figure 2 shows the average time per completed interview over the course of IWS using the stricter definition of completed interviews.[4] As is typical of most surveys, the average interview time initially drops as enumerators become increasingly adept at administering the survey, and then stabilizes (Olson and Smyth, 2020). In IWS, the stabilization took about two weeks: 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. This large variance is a little disconcerting, but, it is precisely what we exploit to improve enumerator performance (explained in section 5).

4.3 Ratio of completed to initiated interviews

A low share of completed interviews in all initiated interviews, suggests futility of efforts by the data collection team. Once initiated, an interview could end up being incomplete for multiple reasons, such as, the respondent withdrew consent, stopped the survey mid-way, was not available during re-visits, and the interview did not meet the minimum time criterion to be counted as completed. The PIs must investigate the underlying causes and accordingly take remedial action. This could be instituting a more effective style of delivering consent, fixing prior appointments in consultation with the respondents, rethinking the re-visit protocol and stopping rule,[5] and sensitizing enumerators on spending adequate time with each respondent. Figure 3 shows the ratio of completed to

[4]For interviews that occurred over multiple sittings, the interview time excludes the time between spells.

[5]Stopping rule refers to the maximum number of attempts to complete an interview before it is closed.

initiated interviews over the course of IWS, again using the stricter definition of completed interviews. The survey started off with a high completion rate of about 95 percent, which decreased steadily to about 70 percent. This declining trend 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 the time.[6] Given that IWS focuses on religious identity, respondents, especially Muslims, were fearful of participating, resulting in a lower completion rate.

We used a dashboard to track these parameters. A dashboard is a one-stop place where key performance indicators can be visualized at a glance. Figures 4 and 5 present screen shots of the IWS dashboard. We designed it using 'Shiny', an open-source 'R'-based package for building web applications. PIs need to decide whether to build their own dashboard from scratch using freeware, or use paid, applications, that come with in-built customizable dashboards. Building a dashboard is a highly specialized skill, so the choice would depend upon the pool of available talent and the budget for paradata monitoring.

5 Paradata Flagging of Deviant Enumerators

We used, what are called flags, to identify enumerators exhibiting deviant practices in the field. Once identified, the flagged enumerator's supervisor would talk to them and provide constructive feedback. The basic idea of a flag (explained below), is not novel (Jans et al., 2013). However, their use to improve data quality in an ongoing survey, is still not widespread, especially in developing countries. Below, we present a detailed exposition of how we designed paradata-based flags.

A flag, in our method, involves comparing within a group of enumerators who faced similar field conditions, and identifying those enumerators (if any), whose performance deviated substantially from the group average. We call each such group of enumerators a 'comparison group'. Restricting comparisons to only enumerators facing similar external

[6]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.

conditions is crucial for the credibility of our method as it ensures that different data generating processes are not mixed together. Once comparisons are restricted in this manner, it possible to interpret the group average as the process average in steady state, and deviations from this average as errant behavior requiring intervention.

It is important to note that a flag is only *suggestive* of a faulty practice, and should not be construed as conclusive evidence of wrongdoing. This is because, while it correctly identifies deviant behavior, it does not go into the reasons for it. It is possible, though unlikely,[7] that the said behavior was the right response given the circumstances on the field. It is imperative that PIs emphasize this aspect to the field supervisors so that their conversations with flagged enumerators are not accusatory in nature. Next, we discuss crucial design features of our method for creating flags.

5.1 Defining a comparison group

It is crucial that a comparison group only consists of enumerators facing similar field conditions, who are then expected to display similar behavior under normal circumstances. The tenet to follow when defining a comparison group is that optimally defined groups would maximize between-group variability and minimize within-group variability under stable field conditions. We defined a comparison group as a specific state (Karnataka/Rajasthan), subregion (urban/rural), and gender (of enumerator) combination, resulting in eight such groups. Below, we explain the rationale for including each delineating dimension.

Consider enumerators operating in the same state and same subregion. Recall that a respondent was to be interviewed by an enumerator of the same gender. It would therefore be incorrect to bracket male and female enumerators into one comparison group as they face very different respondents, because of which we anticipate them to follow different protocols on average. In fact, the IWS questionnaire is itself gender-specific (see Appendix A), leading to different average interview times by enumerator gender. If males and females were grouped together, within-group variability would be high, violating the

[7] The reason it is unlikely is that when flagging errant behaviour we take care to only compare across enumerators facing similar field conditions. Consequently, we expect them to follow similar practices.

basic definition of a comparison group. Following similar reasoning, one can rationalize the use of state and subregion as delineating dimensions.

5.2 Setting performance-window length

Another highlight of our method is what we call *dynamic benchmarking*. Instead of examining enumerator performance in a cumulative fashion, we studied it in separate blocks of time, namely, one week at a time. By doing so, performance was flagged as deviant against a moving benchmark that accounted for all secular changes over time. For instance, as enumerators gain proficiency, there is a secular decline in the average interview duration and separate weekly windows would correctly account for this.

There is a tradeoff when deciding the appropriate window length. If the window is too long, faulty practices would continue unchecked. On the other hand, if it is too short, there may not be enough data points within a comparison group for the underlying statistical theory to operate, invalidating the credibility of our flags. Additionally, shorter windows may dictate more frequent interventions, which is costly in terms of supervisors' time. Looking back at our own experience, a two-week window may have been more effective in managing this trade-off.

5.3 Setting thresholds for errant behavior

How far away from the group mean should a value be in order to be flagged? Some studies have referred to the three-sigma rule, i.e., three standard deviations away from the mean, as a statistical benchmark (Jans et al., 2013). However, even they acknowledge that no single rule fits all.

We used two, somewhat arbitrary, thresholds, namely, 1 and 1.6 standard deviations. In general, thresholds could be set based on a pilot phase, and driven by feasibility considerations. For example, if a particular threshold triggers a large number of flags, in turn requiring a large number of (costly) interventions, it may be prudent to set a limit that would trigger fewer flags.

5.4 Choosing flags

Each flag is associated with a specific field practice that we would like to monitor. Since the marginal cost of designing an additional flag is small, there is a tendency to create too many, not recognizing the flip side. To elaborate, imagine a scenario where a field supervisor has been asked to talk to four flagged enumerators, sounding out each one on six different practices. First, it is likely that much of the crucial feedback would be lost in translation, especially if there are nuanced flags that are hard to talk about. Moreover, being warned on too many fronts increases the cognitive burden on the enumerator. They may feel overwhelmed, diminishing their ability to take remedial action. Worse still, they may feel dejected and give up entirely. While the numbers in this example are arbitrary, we alert PIs to be judicious when choosing flags. Our choice was driven by a focus on data quality rather than survey timeliness as we felt that the agency already had checks on the latter. We were more concerned about enumerators violating interview protocols in order to meet the agency's productivity targets.

Table 2 presents the flags we monitored. In column (3), against each flag, we specify the main performance dimension(s) it evaluates. We consider three dimensions: (a) 'Content knowledge' refers to a sound understanding of the concepts used in the questionnaire; (b) 'Effort exerted' is proxied by the amount of time the enumerator spent with the respondent; and (c) 'Ethics' is about adherence to interview protocols. In columns (4) and (5) we describe each flag in terms of the field practice it monitors, and the underlying concern any deviation gives rise to. The last three columns provide details about flag design and cover the following aspects: (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 method used for ranking enumerators for intervention.

Our list of flags is neither prescriptive nor exhaustive. In fact, if we were to do this again, we would cut down the number of time based flags to only two: 'Survey Time' for the whole interview, and 'Section4 Time', given the significance of section 4 in meeting

IWS objectives. Appendix B provides a rationale for using 'Survey Time' as a catch all for other time based flags.

6 Paradata Based Intervention

We first describe our interventions, and then show that they had an impact on actual enumerator behavior in the field.

6.1 Timeline of interventions

Once generated, the flags were collated into weekly reports, one for each state. Appendix C presents a sample report. The reports were shared with the respective state level manager, who in turn emailed it to all field supervisors. The manager followed this up with a phone conversation with each field supervisor where only information concerning the enumerators under them was discussed. The final step involved a private conversation between the field supervisor and a flagged enumerator. The field supervisor was advised to point out the deviant behavior without being accusatory, and nudge the enumerator to take corrective action as required.

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. 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 some time before they stabilize. The first report was shared on March 4 and March 3 in Karnataka and Rajasthan, respectively; while the second report was shared on March 10 and March 14, respectively.[8] The survey was officially stopped on March 17, but no new interviews were closed after March 14, making it the effective end date.

In subsection 6.2, we examine interventions based on only the first report and disregard the second report. This is because: (a) Just around the time that the second report was shared, Covid-19 was beginning to impact enumerator psyche, and it would

[8]Sharing of the second report was delayed in Rajasthan because of a short pause in field operations on account of *Holi*, a festival mainly celebrated in north India.

not be possible to disentangle the effect of this from that of our interventions; (b) The second report is likely to interact with the first, making it impossible to separate the independent effects of each. (c) There is no post-intervention period for the second report in Rajasthan.

6.2 Effectiveness of interventions

We use Ordinary Least Squares regressions with enumerator fixed effects to analyze the impact of interventions based on the first report. We estimate the following equation:

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

Here, i is for interview, j for enumerator, and k for a specific flag such as Survey Time or Section4 Skip. $Performance$ refers to the particular field practice that a flag monitors. For example, in case of Survey Time, it is the interview duration in minutes; and for Section4 Skip, it is a binary indicator for whether the respondent was reported as 'Not Working'. $FlagSame$ and $FlagOther$ are indicators for whether the enumerator was flagged for flag k, and for some other flag ($\sim k$), 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. ϵ captures all idiosyncratic factors that affect performance.

The primary coefficient of interest is β_1. It captures the change in performance related to a specific practice as a result of 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 they changed behavior relative to an enumerator's own behavior 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. In order to improve the precision of our estimates, we restricted the regressions to enumerators with at least ten completed interviews. Standard errors are clustered at the enumerator level.

Table 3 presents descriptive statistics on flags, along with the 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 6 and March 5 for Karnataka and Rajasthan, respectively. During the analysis period, a total of 88 enumerators completed at least one interview, of which 46 were women. Only those flags are studied for which at least one enumerator was flagged in the report. Column (3) shows the number of enumerators flagged against each flag. Columns (4) and (5) present the mean value of the field practice, during the pre-intervention period, for all enumerators 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, Section4 Time. They increased the interview time for section 4 by 0.7 minutes, amounting to 18 percent of the pre-intervention average time for this section. At the same time, they reduced the interview time for section 8 by 0.2 minutes (Section8 Time). We conjecture that this could be due to (a) enumerators compensating for increased time in some sections by cutting back time spent elsewhere,[9] and (b) too many flags adversely affecting the communication between supervisors and flagged enumerators. Our conjecture is partly strengthened by some significant estimates for β_2. We see that flagging for some other practice increased the time spent on sections 4 and 5 (Section5 Time), and lowered the number of respondents reported as 'Not Working' (Section4 Skip), though the last result is statistically significant only at the 10 percent level. This suggests that intervening made a difference and changed behavior in the field, but, not always along targeted lines.

[9]It is not hard to imagine that being the core of the survey, section 4 was emphasized during feedback sessions, making it likely that enumerators focused much more on it when taking corrective action.

7 Lessons Learnt

Some of what we share below is specific to the IWS context where data collection was outsourced to an external agency, and data was collected using CAPI.

7.1 Understanding the structure and composition of paradata

The way in which raw data is organized varies across projects. In our case, the structure was fairly simple: the unit of observation for both paradata and survey data was an individual respondent. There could be more complex structures, where, one or both datasets is structured differently. For example, the unit of observation could be an episode of interaction with the respondent resulting in multiple entries for some respondents. While we are agnostic about which structure is better, it is important to know beforehand how the raw data would be organized. A good way to accomplish this is to pilot paradata operations along with the main survey.

It is equally important to know the exact paradata variables that will be *shared*. Knowing the granularity of time stamps, whether they are at the interview, section, or question level, is essential when deciding which flags are feasible. To avoid scrutiny of its operations, the external agency may not always be forthcoming in sharing detailed paradata. It is therefore important to dialogue with them right and have them fully on board with all aspects of paradata monitoring. It would be ideal if paradata requirements could be included as deliverables in the formal contract with the agency.

7.2 Ensuring high quality paradata

We discuss quality checks for a few crucial paradata variables.

Time stamps: System settings on all devices should be checked before data collection begins. To avoid tampering, it should not be possible to change dates and times once they have been entered.

Duration data: When certain activities are not clocked between time stamps, such as administration of consent, question-times may not add up to section-time. In order to design

effective flags, it is important to understand what parts of the respondent-interviewer interaction are covered between timestamps, and the hierarchical links between duration variables of varying granularity.

Visit result: At the time of closing a case, an enumerator is required to mark the status of the interview as complete, incomplete, door refusal, or not available. In our experience, enumerators are very often not clear about how to code this correctly. To avoid this, the 'Visit result' variable should be emphasized during training.

Enumerator identifier: It should preferably be a single variable (not a combination of many variables), selected from a drop-down list of enumerator names (not codes). This would eliminate mis-spellings and use of multiple codes for the same enumerator.

7.3 Choosing between dashboard and printed reports

Our advice on this is contrary to the push towards dashboards in recent literature (Sarikaya et al., 2019; Yigitbasioglu and Velcu, 2012). While highly sophisticated dashboards that generate automated reports in real time are undoubtedly preferred over a system that generates manual reports with a lag, they may not always be feasible. Dashboard design requires specialized coding skills, and when these are scarce it may be prudent to use manual reports instead.

In IWS, we found the dashboard very useful for tracking overall progress of the survey, but we did not use it for the flagging exercise. Instead, we relied on paper reports, created using substantial manual intervention. A dashboard is merely a tool, and if, other, more cost-effective tools are available, dashboard design should not be considered a pre-requisite for paradata-based interventions.

7.4 Principles of dashboard/report design

Mohadjer and Edwards (2018) present a detailed account of dashboard design. We highlight two design issues from their work. First, a dashboard view should be designed keeping only one type of end-user in mind. Targeting multiple users at once would make it harder for each type to access the information they need. For household surveys, two

views could be created, one for the PIs to monitor survey progress, and another for the field supervisors to track their own enumerators. Second, less may be better than more when it comes to dashboard design. Using a layered design that highlights a few salient aspects in the first view, with inner views providing a limited number of necessary details, is better than a flat design which displays too many moving parts. This advice is equally applicable when printed reports substitute dashboards.

In IWS, the basis for intervention was a weekly report. As seen in Appendix C, our report is organized flag-wise, listing the names of flagged enumerators under each flag. Given that the unit of intervention is an enumerator, it would have been more effective to organize it enumerator-wise. This would immediately clarify who are the flagged enumerators and the specific practices to review with each one.

7.5 Frequency of intervention

Intervening to change enumerator behavior is costly in terms of the field supervisors' time. Not only do frequent interventions increase supervisors' workload, but also leave insufficient time for enumerators to introspect and take corrective action. It is important that the PIs deliberate on the intervention process in its entirety to make it most effective for them.

8 Discussion

We presented a complete prototype of how to operationalize paradata use during an ongoing survey, in a relatively resource constrained environment. We found dashboard tracking of three specific paradata based parameters to be extremely effective in monitoring survey progress. In terms of using paradata based flags to streamline enumerator practices, ex-post regression analyses suggests that we had some success in influencing enumerator behavior in the desired direction. We hope that our work encourages data collectors, especially those from developing countries, to harness paradata to improve their surveys.

While we have clarified some aspects of paradata use to reduce interviewer generated survey error, we have left out some important forms of paradata discussed elsewhere in the literature. These include GPS coordinates (Bhuiyan and Lackie, 2016; Edwards et al., 2017; Montalvo et al., 2018), audio recordings of interviews (Bhuiyan and Lackie, 2016; Gomila et al., 2017; Hicks et al., 2010), and interviewer observations (West and Kreuter, 2018).

In their comprehensive review of survey research organizations, Murphy et al. (2016) are unable to identify a single set of best practices that could serve as a model when trying to mitigate interviewer effects. Cohen and Warner (2021) are among the first to address this gap in the literature. They present systematic evidence on the relative merits of different quality control procedures, including that of paradata generated flags. More such work is needed to arrive at a standardized set of procedures, and to eliminate redundancy across multiple methods and indicators used within a single survey.

Paradata has tremendous potential to improve survey quality which remains under-utilized, especially in low and middle income countries. One way to encourage its use is for donor agencies that fund surveys to: a) mandate paradata use, b) provide a budget specifically earmarked for it, and c) require that some paradata be made public along with the main survey data.[10] This could make paradata use a standard practice world-wide. Profit oriented data collection agencies may then begin to view paradata not as a threat to their commercial interests, but as an integral tool to improve their business.

[10]Interview length is a good example of such a paradata item.

References

Ackermann-Piek, D., J. M. Korbmacher, and U. Krieger (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, and B. T. West (Eds.), *Interviewer Effects from a Total Survey Error Perspective* (1 ed.)., Chapter 14, pp. 193–206. Boca Raton: Chapman and Hall/CRC.

Bhuiyan, M. F. and P. Lackie (2016). Mitigating survey fraud and human error: Lessons learned from a low budget village census in bangladesh. *IASSIST Quarterly 40*(3), 20–26.

Bushery, J. M., J. W. Reichert, K. A. Albright, and J. C. Rossiter (1999). Using date and time stamps to detect interviewer falsifcation. *Proceedings of the survey research method section*, 316–20.

Choumert-Nkolo, J., H. Cust, and C. Taylor (2019). Using paradata to collect better survey data: Evidence from a household survey in tanzania. *Review of Development Economics 23*(2), 598–618.

Cohen, M. J. and Z. Warner (2021). How to get better survey data more efficiently. *Political Analysis 29*(2), 121–138.

Couper, M. (1998). Measuring survey quality in a casic environment. *Proceedings of the Survey Research Methods Section of the ASA at JSM 1998*, 41–49.

Da Silva, D. N. and C. J. Skinner (2020). Testing for measurement error in survey data analysis using paradata. *Biometrika 108*(1), 239–246.

Edwards, B., A. Maitland, and S. Connor (2017). Measurement error in survey operations management. In *Total Survey Error in Practice*, Chapter 12, pp. 253–277. John Wiley & Sons, Ltd.

Edwards, B., H. Sun, and R. Hubbard (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* (1 ed.)., Chapter 6, pp. 77–89. Boca Raton: Chapman and Hall/CRC.

Finn, A. and V. Ranchhod (2015, 09). Genuine Fakes: The Prevalence and Implications of Data Fabrication in a Large South African Survey. *The World Bank Economic Review 31*(1), 129–157.

Gomila, R., R. Littman, G. Blair, and E. L. Paluck (2017). The audio check: A method for improving data quality and detecting data fabrication. *Social Psychological and Personality Science 8*(4), 424–433.

Guyer, H. M., B. T. West, and W. Chang (2021). The interviewer performance profile (ipp): A paradatadriven tool for monitoring and managing interviewer performance. *Survey Methods: Insights from the Field. Retrieved from https://surveyinsights.org/?p=15306*.

Hicks, W. D., B. Edwards, K. Tourangeau, B. McBride, L. D. Harris-Kojetin, and A. J. Moss (2010, 01). Using cari tools to understand measurement error. *Public Opinion Quarterly 74*(5), 985–1003.

Hood, C. C. and J. M. Bushery (1997). Getting more bang from the reinterview buck: Identifying "at risk" interviewers. *Proceedings of the survey research method section*, 820–24.

Jans, M., S. Sirkis, and D. Morgan (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*, Chapter 9, pp. 191–229. Hoboken, NJ: Wiley.

Kosyakova, Y., L. Olbrich, J. Sakshaug, and S. Schwanhäuser (2019). Identifcation of

interviewer falsifcation in the iab-bamf-soep survey of refugees in germany. *FDZ-METHODENREPORT: Methodological aspects of labour market data*.

Kreuter, F., M. Couper, and L. Lyberg (2010). The use of paradata to monitor and manage survey data collection. *Section on Survey Research Methods – JSM 2010*, 283–296.

Krueger, B. S. and B. T. West (2014). Assessing the Potential of Paradata and Other Auxiliary Data for Nonresponse Adjustments. *Public Opinion Quarterly 78*(4), 795–831.

Krueter, F. and K. Olson (2013). Paradata for nonresponse error investigation. In F. Krueter (Ed.), *Improving Surveys with Paradata: Analytic Uses of Process Information*, Chapter 2, pp. 13–42. Hoboken, NJ: Wiley.

Kuriakose, N. and M. Robbins (2018). Don't get duped: Fraud through duplication in public opinion surveys. *Statistical Journal of the IAOS 21*(1), 195–214.

Lupu, N. and K. Michelitch (2018). Advances in survey methods for the developing world. *Annual Review of Political Science 32*, 283–291.

Mohadjer, L. and B. Edwards (2018). Paradata and dashboards in piaac. *Quality Assurance in Education 26*(2), 263–277.

Montalvo, J. D., M. A. Seligson, and E. J. Zechmeister (2018). Improving adherence to area probability sample designs: Using lapop's remote interview geo-locating of households in real-time (right) system. *Americas Barometer Methodological Note IMN004*.

Murphy, J., P. Biemer, C. Stringer, R. Thissen, O. Day, and Y. P. Hsieh (2016). Interviewer falsification: Current and best practices for prevention, detection, and mitigation. *Statistical Journal of the IAOS 32*, 313–326.

Olson, K. and J. D. Smyth (2020). What Do Interviewers Learn?: Changes in Interview Length and Interviewer Behaviors over the Field Period. 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 (1 ed.)., Chapter 20, pp. 279–291. Boca Raton: Chapman and Hall/CRC.

Olson, K., J. D. Smyth, J. Dykema, A. L. Holbrook, F. Kreuter, and B. T. West (Eds.) (2020). *Interviewer Effects from a Total Survey Error Perspective*. Chapman and Hall/CRC.

Pashazadeh, F., A. Cernat, and J. W. Sakshaug (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, and B. T. West (Eds.), *Interviewer Effects from a Total Survey Error Perspective* (1 ed.)., Chapter 16, pp. 221–234. Boca Raton: Chapman and Hall/CRC.

Sarikaya, A., M. Correll, L. Bartram, M. Tory, and D. Fisher (2019). What do we talk about when we talk about dashboards? *IEEE Transactions on Visualization and Computer Graphics 25*(1), 682–692.

Schaeffer, N. C., J. Dykema, and D. W. Maynard (2010). Interviewers and interviewing. In P. V. Marsden and J. D. Wright (Eds.), *Handbook of survey research*, Chapter 13, pp. 437–470. Bingley, UK: Emerald Group Publishing.

West, B. T. and A. G. Blom (2016). Explaining Interviewer Effects: A Research Synthesis. *Journal of Survey Statistics and Methodology 5*(2), 175–211.

West, B. T. and R. M. Groves (2013). A Propensity-Adjusted Interviewer Performance Indicator. *Public Opinion Quarterly 77*(1), 352–374.

West, B. T. and F. Kreuter (2018). Strategies for increasing the accuracy of interviewer observations of respondent features: Evidence from the u.s. national survey of family growth. *Methodology (Gott) 14*(1), 16–29.

Yan, T. and K. Olson (2013). Paradata for nonresponse error investigation. In F. Krueter (Ed.), *Analyzing paradata to investigate measurement error*, Chapter 4, pp. 73–95. Hoboken, NJ: Wiley.

Yigitbasioglu, O. M. and O. Velcu (2012). A review of dashboards in performance management: Implications for design and research. *International Journal of Accounting Information Systems 13*(1), 41–59.

Tables

Table 1: 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

Table 2

Paradata Flags to Monitor Enumerator Performance (1/6)							
S. no.	Flag name	Dimension being evaluated	Field practice being monitored	Underlying concern	Based on paradata or survey data	Flag 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 resulting in poor quality data.	Paradata	An 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-gender strata, top three enumerators with highest positive ratios (if any), were flagged for intervention.
2	Section0 Time	Effort	Time taken to field the Household Register section.	Very short section time suggests violation of interview protocols resulting in poor quality data.	Paradata	An 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-gender strata, top three enumerators with highest positive ratios (if any), were flagged for intervention.
3	Section1 Time	Effort	Time taken to field the Demographic Characteristics section.	Very short section time suggests violation of interview protocols resulting in poor quality data.	Paradata	An 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-gender strata, top three enumerators with highest positive ratios (if any), were flagged for intervention.

Para data Flags to Monitor Enumerator Performance (2/6)							
S. no.	Flag name	Dimension being evaluated	Field practice being monitored	Underlying concern	Based on paradata or survey data	Flag threshold	Ranking enumerators for intervention
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
4	Section2 Time	Effort	Time taken to field the Household Living Standards section.	Very short section time suggests violation of interview protocols resulting in poor quality data.	Paradata	An 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-gender strata, top three enumerators with highest positive ratios (if any), were flagged for intervention.
5	Section3 Time	Effort	Time taken to field the Activity Profile for the Last Year section	Very short section time suggests violation of interview protocols resulting in poor quality data.	Paradata	An 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-gender strata, top three enumerators with highest positive ratios (if any), were flagged for intervention.
6	Section4 Time	Effort	Time taken to field the Weekly Labour Force Status section.	Very short section time suggests violation of interview protocols resulting in poor quality data.	Paradata	An 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-gender strata, top three enumerators with highest positive ratios (if any), were flagged for intervention.

Para data Flags to Monitor Enumerator Performance (3/6)							
S. no.	Flag name	Dimension being evaluated	Field practice being monitored	Underlying concern	Based on paradata or survey data	Flag threshold	Ranking enumerators for intervention
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
7	Section5 Time	Effort	Time taken to field the Household Production Activities section.	Very short section time suggests violation of interview protocols resulting in poor quality data.	Paradata	An 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-gender strata, top three enumerators with highest positive ratios (if any), were flagged for intervention.
8	Section8 Time	Effort	Time taken to field the Decision Making section.	Very short section time suggests violation of interview protocols resulting in poor quality data.	Paradata	An 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-gender strata, top three enumerators with highest positive ratios (if any), were flagged for intervention.
9	Section9 Time	Effort	Time taken to field the Intergenerational Mobility section.	Very short section time suggests violation of interview protocols resulting in poor quality data.	Paradata	An 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-gender strata, top three enumerators with highest positive ratios (if any), were flagged for intervention.

Para data Flags to Monitor Enumerator Performance (4/6)							
S. no.	Flag name	Dimension being evaluated	Field practice being monitored	Underlying concern	Based on paradata or survey data	Flag threshold	Ranking enumerators for intervention
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
10	Section10 Time	Effort	Time taken to field the Social Networks section.	Very short section time suggests violation of interview protocols resulting in poor quality data.	Paradata	An 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-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 an enumerator deliberately leaves out some household members, is may adversely affect survey representativeness.	Survey data	An 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 the respondent's social network.	If an enumerator does not capture everyone in the respondent's network, it may bias analyses based on network structure.	Survey data	An enumerator got flagged when their standardized average network size is below -1.[3]	Intervene on all flagged enumerators (if any).

Para data Flags to Monitor Enumerator Performance (5/6)							
S. no.	Flag name	Dimension being evaluated	Field practice being monitored	Underlying concern	Based on paradata or survey data	Flag threshold	Ranking enumerators for intervention
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
13	Odd Start	Ethics	Whether interview started outside the usual survey hours.	Indicates interview falsification.	Paradata	An 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.
14	Alone Section7	Ethics	Whether an enumerator reports that the respondent was interviewed in private when fielding the Discrimination section.	If an enumerator reports being always alone or being never alone, it is likely that they are not documenting the privacy status correctly, rendering this information useless for analyses.	Survey data	An enumerator got flagged if for all completed interviews in that week they recorded either being always alone or being never alone with the respondent.	Intervene on all flagged enumerators (if any).
15	Section4 Skip	Content, Effort, Ethics	Whether respondent's work status is reported as 'Not Working'.	If the enumerator is not clear about what constitutes work, or does not probe enough, or deliberately records 'not working' to avoid subsequent sections, it may lead to biased estimates of work status.	Survey data	An interview got flagged if the respondent was reported as 'not working'.	For each enumerator, ratio of flagged interviews to completed interviews, in that week, was calculated. Within each state-gender strata, top three enumerators with highest positive ratios (if any), were flagged for intervention.[4]

Notes: Para data Flags to Monitor Enumerator Performance (6/6)
[1]The IWS questionnaire has 13 sections. A section got 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. Using this criterion, the sections that got 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.
[4]Effectively, for this flag, the comparison group is state-gender and not state-region-gender.

Table 3

Effect of Para Data based Interventions on Enumerator Performance (1/4)									
		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	No. 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	Section2 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	Section4 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	Section5 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)			

Effect of Para Data based Interventions on Enumerator Performance (2/4)									
		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	No. of Observations/ Completed Interviews (Number of Clusters/ Enumerators)
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
5	Section8 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	Section9 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	Section10 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)			
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)			

Effect of Para Data based Interventions on Enumerator Performance (3/4)									
		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	No. of Observations/ Completed Interviews (Number of Clusters/ Enumerators)
[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
9	Alone Section7 (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	Section4 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)			

Notes: Effect of Para Data based Interventions on Enumerator Performance (4/4)

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. Section0 Time, Section1 Time, Section3 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 regression analysis is based on enumerator performance between February 17 and March 10 for Karnataka, and between February 17 and March 14 for Rajasthan.
The descriptive statistics are limited to enumerators who completed at least 1 interview (by the strict definition of a completed interview), while the regressions are limited to enumerators who completed at least 10 interviews.
Regressions are at the interview level, and the dependent variable is indicated under the column (2) Flag Name.
Standard deviations (for descriptive statistics) / Clustered standard errors (for regression coefficients) are shown in parentheses.
* stands for statistical significance at the 10 percent level of significance, ** at 5 percent, and *** at 1 percent.

[1]Section2 was only administered 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 till the date of intervention (March 4 for Rajasthan and March 5 in Karnataka).

Figures

Figure 1

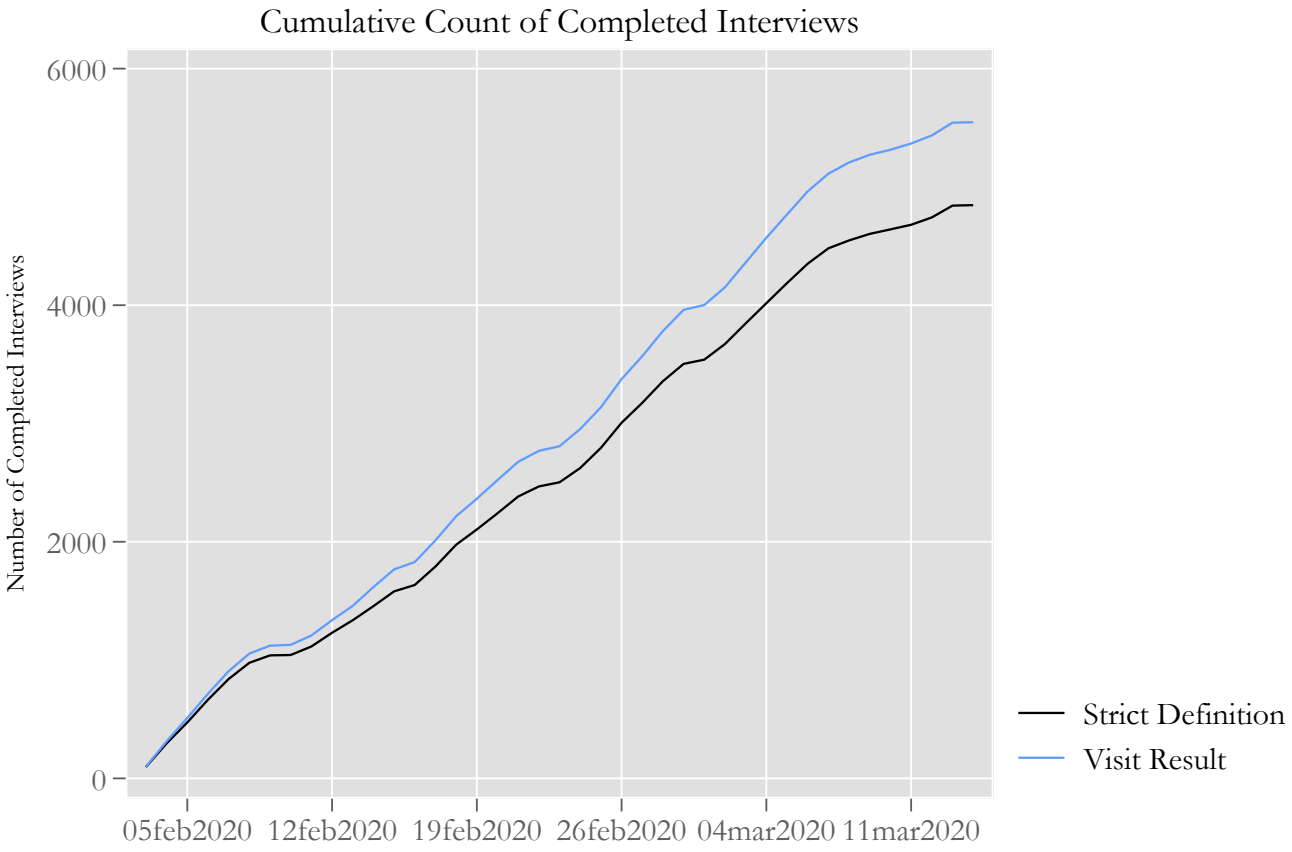


Figure 2

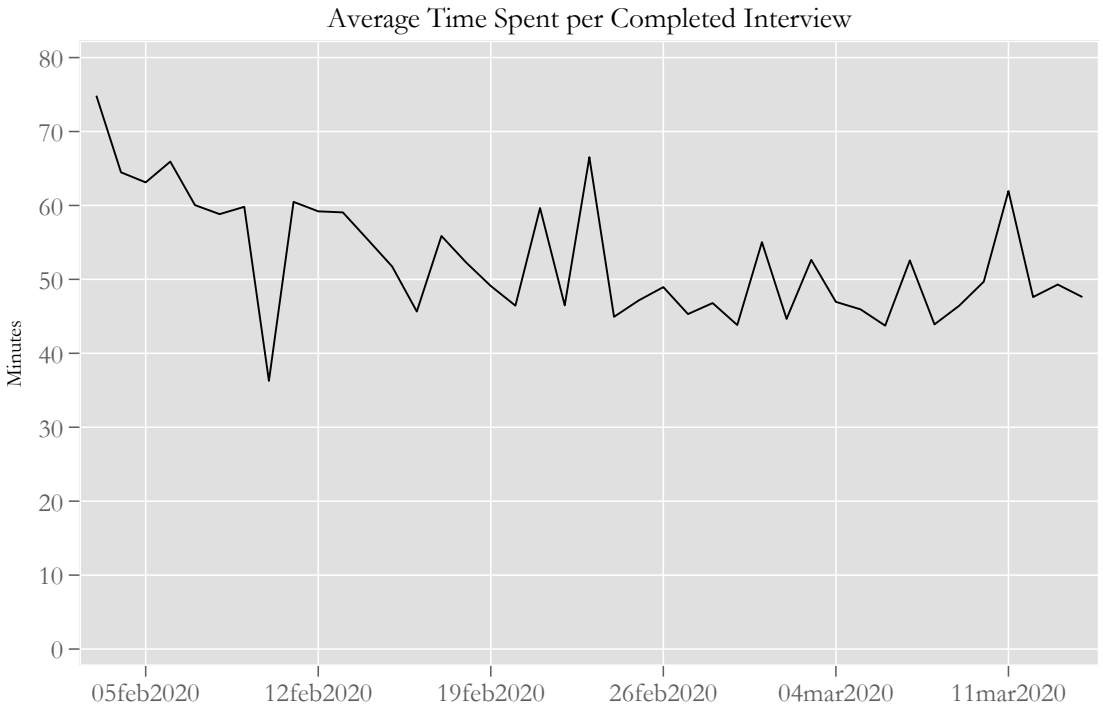


Figure 3

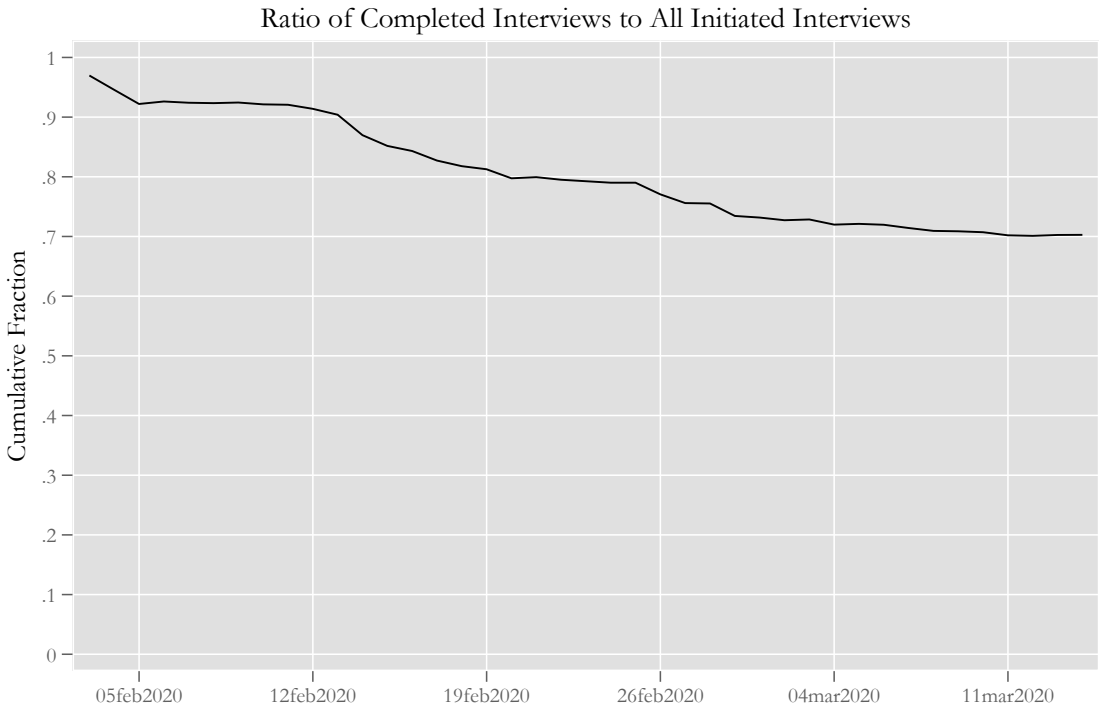


Figure 4: First View

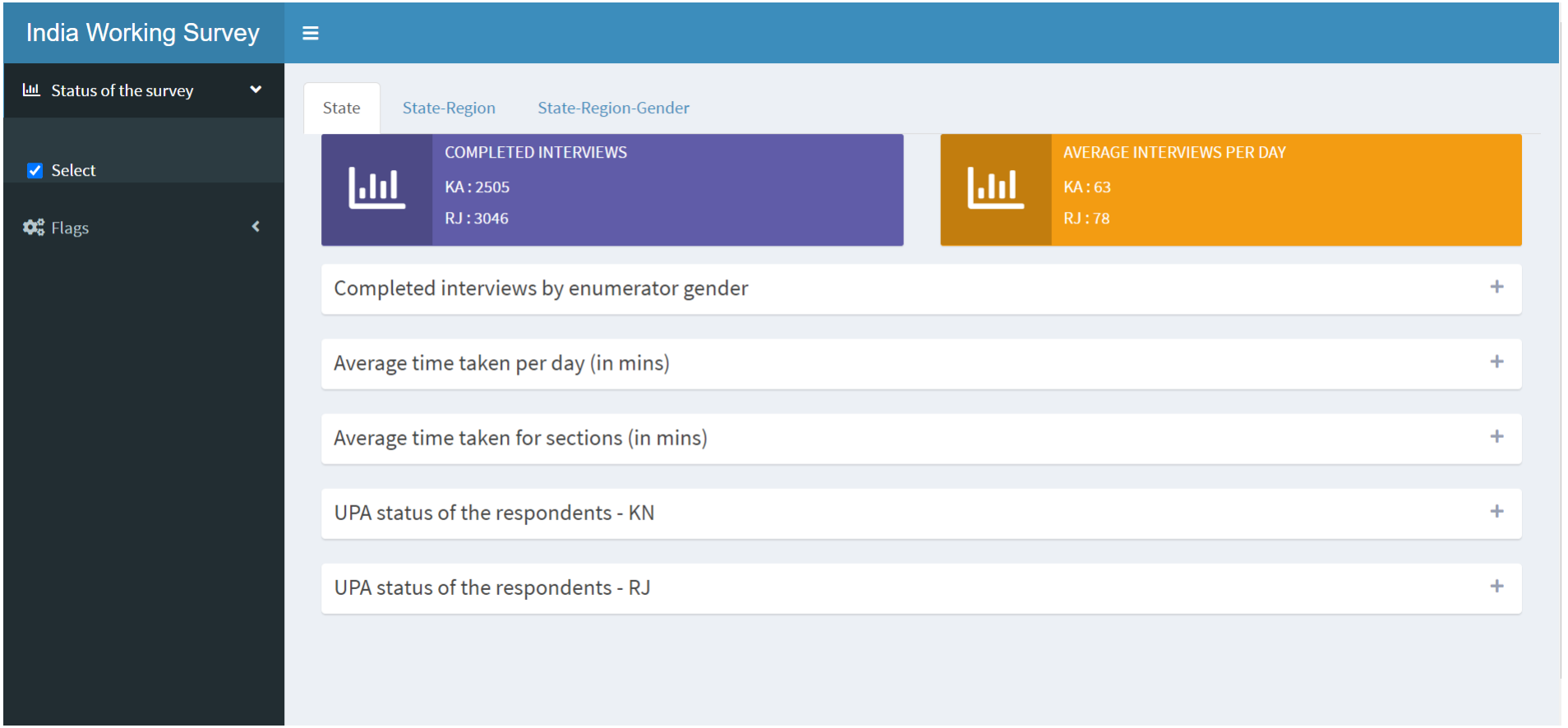
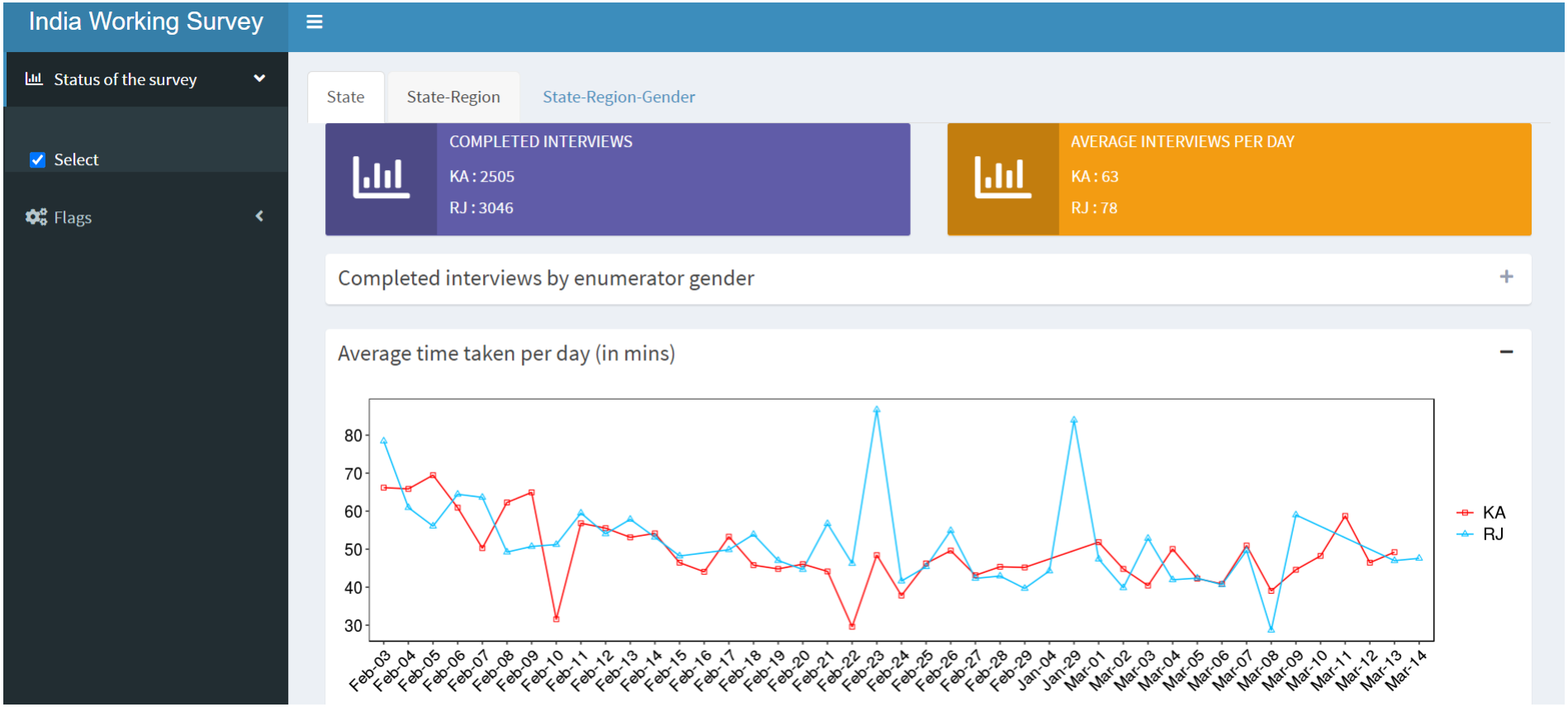


Figure 5: Inner View



Appendices

A Organization of IWS Field Questionnaire

The table below presents the section-wise organization of the IWS field questionnaire. The columns labeled Female and Male, show the approximate count of questions in each section fielded by the female and male enumerators, respectively.

IWS Field Questionnaire			
Section	Description	Female	Male
0. Household Register	Basic household roster. Fielded only by the female enumerator.	6	
1. Demographic Characteristics	Demographic information such as caste, education, and major work status. The female enumerators recorded it for all household members, the male enumerator for only the male respondent.	22	17
2. Household Living Standards	Information about the dwelling, household amenities, and assets. Fielded only by the female enumerator.	12	
3. Activity Profile for the Last Year	Major work activity status and skill of the respondent.	26	26
4. Weekly Labour Force Status	Detailed information on respondent's activities in the week prior to the interview.	58	58
5. Household Production Activities	Time spent by the respondent on household production activities in 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 they extend.	5	5
11. Women Out of Work Force	Information about women respondents who reported their major work status to be 'not working'.	9	9
12. Students	Information about respondents who reported their major work status to be 'studying or attending an education institution'.	2	2
13. Unemployed	Information about respondent who reported their major work status to be 'unemployed'.	9	9

B Relationship between Survey Time and Other Section Times

The motivation for this appendix is to examine whether including the 'Survey Time' flag allows us to do away with other flags based on individual section times. Here, we only include those sections for which at least one enumerator was flagged in the first paradata based report. These are sections 2, 4, 5, 8, 9, and 10. An important caveat is that we only have a few data points, and so this analysis should not be taken as conclusive evidence.

In the first report, spanning the period between February 17 and February 23, a total of 12 enumerators were flagged for Survey Time. In terms of completed interviews, 16 percent of the 868 completed interviews were flagged for Survey Time.[11] The table below presents corresponding numbers for time flags based on other sections, and also shows the overlap, if any, between Survey Time and the other section-time flags.

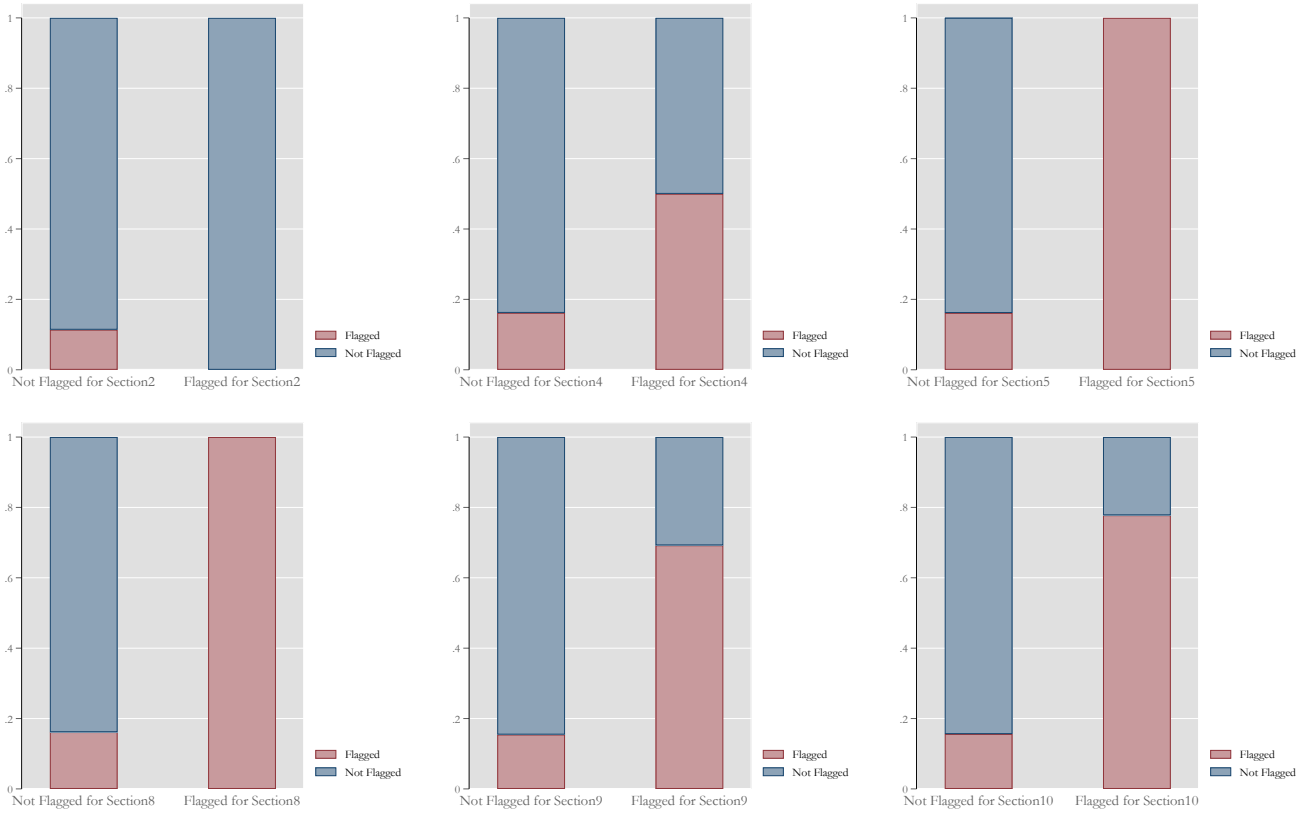
Overlap between Survey Time Flag and Other Section Time Flags				
Flag Name	Number of Enumerators Flagged	Number of Flagged Enumerators in Common with Survey Time	Number of Interviews Flagged	Number of Flagged Interviews in Common with Survey Time
Survey Time	12		141	
Section2 Time	5	0	5	0
Section4 Time	2	0	2	1
Section5 Time	1	1	1	1
Section8 Time	1	0	1	1
Section9 Time	6	3	13	9
Section10 Time	7	3	9	7

The figure below shows the status (flagged or not) of Survey Time flag conditional on the status of each of the other section time flags. If Survey Time is to act as a good stand-in for the other section time flags, then conditional on being flagged for a particular section, the interview should also be flagged for Survey Time. In other words, in each

[11] These include flagged interviews of enumerators who may not be among the 12 flagged enumerators. This is because, in spite of having some flagged interviews, an enumerator may not get flagged if their ratio of flagged to completed interviews is not among the top three.

sub-plot below, the red shaded area in the second bar should be large. Except for Section 2, Survey Time does a pretty good job in this respect.

Survey Time Flag Status Conditional on Flag Status of Other Section Times



C Sample Report Used for Intervention in IWS

Enumerator Report for Week: 17th February through 23rd February

Find below the list of flagged enumerators. They have been flagged because as compared to the other enumerators they are doing something very different in some of their interviews. It is, therefore, important for the supervisors to talk to them and figure out why this is the case. The flagged enumerators may not necessarily be doing something wrong. **It is important that the supervisors do not assign blame when talking to enumerators.**

Below the list of flagged enumerators, you will also find information about the particular interviews for which the enumerator is being flagged. For the first flag, that is, Survey Time, the data on flagged interviews is given in a separate Excel file, but for all other flags, the flagged interviews are shared in this report itself. The supervisors may want to use the information about specific interviews when talking to enumerators.

1) Survey Time: These enumerators are taking less time to complete some important sections in the survey. The sections being tracked are— Section 1: Demographic Characteristics, Section 5: Household Production Activities, Section 7A: Discrimination, Section 8: Decision Making, and Section 10: Networks. The concern here is that the flagged enumerators are going through the survey very fast, and this may result in poor data quality.

(Average duration per interview for female enumerators is 14.42 minutes.)
(Average duration per interview for male enumerators is 13.95 minutes.)

In the table below, the value in parentheses against each enumerator shows the share of completed interviews by that enumerator that took very little time.

Gender	Karnataka
Female(s)	Ms. X1 (0.37)
	Ms. X2 (0.33)
	Ms. X3 (0.3)
Males(s)	Mr. Y1 (0.5)
	Mr. Y2 (0.42)
	Mr. Y3 (0.33)

The flagged interviews against each enumerator are given in a separate Excel file.

2) Section0 Time: These enumerators are taking less time to complete Section 0, 'The Household Register'. The concern here is that the flagged enumerators are going through the section very fast, taking very short times for their interviews, and this may result in poor data quality. This flag is only applicable for female enumerators, as male enumerators do not field this section.

(Average duration per interview for this section for female enumerators is 3.5 minutes.)
(Average duration per interview for this section for male enumerators is 3.3 minutes.)

In the table below, the value in parentheses against each enumerator shows the share of completed interviews by the enumerator that had short section time.

Gender	Karnataka
Female(s)	

3) Section1 Time: These enumerators are taking less time to complete Section 1, 'Demographic Characteristics'. The concern here is that the flagged enumerators are going through the section very fast, taking very short times for their interviews, and this may result in poor data quality.

(Average duration per interview for this section for female enumerators is 3.7 minutes.)
(Average duration per interview for this section for male enumerators is 2.1 minutes.)

In the table below, the value in parentheses against each enumerator shows the share of completed interviews by the enumerator that had short section time.

Gender	Karnataka
Female(s)	
Males(s)	

4) Section2 Time: These enumerators are taking less time to complete Section 2, 'Household Living Standards'. The concern here is that the flagged enumerators are going through the section very fast, taking very short times for their interviews, and this may result in poor data quality. This flag is only applicable for female enumerators, as male enumerators do not field this section.

(Average duration per interview for this section for female enumerators is 2.1 minutes.)

In the table below, the value in parentheses against each enumerator shows the share of completed interviews by the enumerator that had a short section time.

Gender	Karnataka
Female(s)	Ms. X1 (0.25)
	Ms. X2 (0.16)

The flagged interviews against each female enumerator are given below.

Surveyor Name	Start Time	End Time	District	Village	Hhld.	Respond. Name	Section 2 Duration secs.
XXX	18 Feb. 13:40:29	19 Feb. 21:54:15	XXX	XXX	XXX	XXX	54
XXX	18 Feb. 10:38:11	18 Feb. 11:14:50	XXX	XXX	XXX	XXX	52

5) Section3 Time: These enumerators are taking less time to complete Section 3, 'Activity Profile for the Last Year'. The concern here is that the flagged enumerators are going through the section very fast, taking very short times for their interviews, and this may result in poor data quality.

(Average duration per interview for this section for female enumerators is 2.5 minutes.)
(Average duration per interview for this section for male enumerators is 3.2 minutes.)

In the table below, the value in parentheses against each enumerator shows the share of completed interviews by the enumerator that had a short section time.

Gender	Karnataka
Female(s)	
Males(s)	

6) Section4 Time: These enumerators are taking less time to complete Section 4, 'Weekly Labour Force Status'. The concern here is that the flagged enumerators are going through the section very fast, taking very short times for their interviews, and this may result in poor data quality.

(Average duration per interview for this section for female enumerators is 3.4 minutes.)
(Average duration per interview for this section for male enumerators is 4.9 minutes.)

In the table below, the value in parentheses against each enumerator shows the share of completed interviews by the enumerator that had a short section time.

Gender	Karnataka
Female(s)	Ms. X1 (0.14)
Males(s)	Mr. Y1 (0.5)

The flagged interviews against each enumerator are given below.

Surveyor Name	Start Time	End Time	District	Village	Hhld.	Respond. Name	Section 4 Duration secs.
XXX	15 Feb. 11:08:36	17 Feb. 20:07:07	XXX	XXX	XXX	XXX	41
XXX	18 Feb. 13:05:12	18 Feb. 18:09:32	XXX	XXX	XXX	XXX	22

7) Section5 Time: These enumerators are taking less time to complete Section 5, 'Time Spent on Household Production Activities'. The concern here is that the flagged enumerators are going through the section very fast, taking very short times for their interviews, and this may result in poor data quality.

(Average duration per interview for this section for female enumerators is 2.0 minutes.)
(Average duration per interview for this section for male enumerators is 2.0 minutes.)

In the table below, the value in parentheses against each enumerator shows the share of completed interviews by the enumerator that had a short section time.

Gender	Karnataka
Female(s)	Ms. X1 (0.1)
Males(s)	

The flagged interviews against each enumerator are given below.

Surveyor Name	Start Time	End Time	District	Village	Hhld.	Respond. Name	Section 5 Duration secs.
XXX	4 Feb. 13:56:37	22 Feb. 11:46:46	XXX	XXX	XXX	XXX	29

8) Section8 Time: These enumerators are taking less time to complete section 8, 'Decision Making'. The concern here is that the flagged enumerators are going through the section very fast, taking very short times for their interviews, and this may result in poor data quality.

(Average duration per interview for this section for female enumerators is 1.7 minutes.)
(Average duration per interview for this section for male enumerators is 1.5 minutes.)

In the table below, the value in parentheses against each enumerator shows the share of completed interviews by the enumerator that had a short section time.

Gender	Karnataka
Female(s)	Ms. X1 (0.125)
Males(s)	

The flagged interviews against each enumerator are given below.

Surveyor Name	Start Time	End Time	District	Village	Hhld.	Respond. Name	Section 8 Duration secs.
XXX	18 Feb. 11:17:55	18 Feb. 21:31:07	XXX	XXX	XXX	XXX	0

9) Section9 Time: These enumerators are taking less time to complete Section 9, 'Intergenerational Mobility'. The concern here is that the flagged enumerators are going through the section very fast, taking very short times for their interviews, and this may result in poor data quality.

(Average duration per interview for this section for female enumerators is 2.0 minutes.)
(Average duration per interview for this section for male enumerators is 2.2 minutes.)

In the table below, the value in parentheses against each enumerator shows the share of completed interviews by the enumerator that had a short section time.

Gender	Karnataka
Female(s)	Ms. X1 (0.4)
	Ms. X2 (0.33)
	Ms. X3 (0.25)
Males(s)	Mr. Y1 (0.25)
	Mr. Y2 (0.08)
	Ms. X1 (0.4)

The flagged interviews against each enumerator are given below.

Surveyor Name	Start Time	End Time	District	Village	Hhld.	Respond. Name	Section 9 Duration secs.
XXX	18 Feb. 12:07:13	18 Feb. 17:28:41	XXX	XXX	XXX	XXX	50
XXX	4 Feb. 13:56:37	22 Feb. 11:46:46	XXX	XXX	XXX	XXX	14
XXX	6 Feb. 12:22:22	22 Feb. 11:40:02	XXX	XXX	XXX	XXX	13
XXX	18 Feb. 11:17:55	18 Feb. 21:31:07	XXX	XXX	XXX	XXX	0
XXX	15 Feb. 13:51:44	17 Feb. 09:00:33	XXX	XXX	XXX	XXX	37
XXX	3 Feb. 11:41:02	22 Feb. 20:58:19	XXX	XXX	XXX	XXX	18
XXX	3 Feb. 15:40:03	23 Feb. 18:23:11	XXX	XXX	XXX	XXX	21
XXX	4 Feb. 10:35:49	23 Feb. 18:12:51	XXX	XXX	XXX	XXX	27
XXX	5 Feb. 12:03:23	23 Feb. 18:31:43	XXX	XXX	XXX	XXX	15
XXX	19 Feb. 9:21:30	19 Feb. 22:30:31	XXX	XXX	XXX	XXX	54
XXX	7 Feb. 16:05:01	23 Feb. 12:35:45	XXX	XXX	XXX	XXX	32
XXX	7 Feb. 12:11:35	23 Feb. 19:48:12	XXX	XXX	XXX	XXX	0
XXX	18 Feb. 12:07:13	18 Feb. 17:28:41	XXX	XXX	XXX	XXX	50

10) Section10 Time: These enumerators are taking less time to complete Section 10, 'Social Networks'. The concern here is that the flagged enumerators are going through the section very fast, taking very short times for their interviews, and this may result in poor data quality.

(Average duration per interview for this section for female enumerators is 3.0 minutes.)
(Average duration per interview for this section for male enumerators is 4.4 minutes.)

In the table below, the value in parentheses against each enumerator shows the share of completed interviews by the enumerator that had a short section time.

Gender	Karnataka
Female(s)	Ms. X1 (0.2)
	Ms. X2 (0.125)
Males(s)	Mr. Y1 (0.16)
	Mr. Y2 (0.125)
	Ms. X1 (0.2)

The flagged interviews against each enumerator are given below.

Surveyor Name	Start Time	End Time	District	Village	Hhld.	Respond. Name	Section 10 Duration secs.
XXX	4 Feb. 13:56:37	22 Feb. 11:46:46	XXX	XXX	XXX	XXX	0
XXX	6 Feb. 12:22:22	22 Feb. 11:40:02	XXX	XXX	XXX	XXX	0
XXX	18 Feb. 11:17:55	18 Feb. 21:31:07	XXX	XXX	XXX	XXX	0
XXX	15 Feb. 9:55:27	17 Feb. 15:19:31	XXX	XXX	XXX	XXX	35
XXX	15 Feb. 11:23:03	17 Feb. 14:24:21	XXX	XXX	XXX	XXX	44

11) Roster Size: These enumerators are recording a smaller number of individuals within a household as eligible for interview. The concern is that they may be deliberately leaving out some eligible adults and only noting those who are available at the time of the first visit.

(Average roster size per interview for female enumerators is 4.43 members.)
(Average roster size per interview for male enumerators is 4.38 members.)

Gender	Karnataka
Female(s)	
Males(s)	

12) Network Size: These enumerators are recording a smaller network size of the main respondent, that is, they are recording that the respondent knows very few people. The concern is that they are not probing enough to get the full network of the respondent.

(Average network size per interview for female enumerators is 2.04 persons.)
(Average network size per interview for male enumerators is 2.60 persons.)

Gender	Karnataka
Female(s)	Ms. X1
Males(s)	

13) Odd Start Time: These enumerators have been flagged because they are reporting odd start times (between 9 pm and 6 am) for some of their interviews.

Gender	Karnataka
Female(s)	Ms. X1
Males(s)	

14) Alone Section7: These enumerators are either reporting that they are 'Always Alone' or are 'Never Alone' with the main respondent for all their interviews. This does not sound truthful, as one would expect some variation in being able to find the respondent all alone when asking questions in Section 7 Discrimination. It is important to stress to the enumerators that they should note the true environment in which they interviewed the respondent when asking questions in Section 7.

If the enumerator has 1, the enumerator is reporting that they are 'Always Alone' with the main respondents for all their interviews. If 0, it means that they are reporting that they are 'Never Alone'.

Gender	Karnataka
Female(s)	Ms. X1 (1)
	Ms. X2 (1)
	Ms. X3 (1)
	Ms. X4 (1)
	Ms. X5 (1)
	Ms. X6 (1)
	Ms. X7 (1)
	Ms. X8 (1)
	Ms. X9 (0)
Males(s)	Mr. Y1 (1)
	Mr. Y2 (1)
	Mr. Y3 (1)
	Mr. Y4 (1)
	Mr. Y5 (1)
	Mr. Y6 (1)
	Mr. Y7 (1)
	Mr. Y8 (1)
	Mr. Y9 (1)
	Mr. Y10 (1)
	Mr. Y11 (0)

15) Section4 Skip: In Section 4 on 'Weekly Labour Force Status', these enumerators are recording that in the last week, the main respondent was not engaged in any work activity, i.e., the enumerator marked the respondent as Not Working. The concern is that they are either not probing enough about work or are recording this so as to skip other questions related to work.

In the table below, the value in parentheses against each enumerator shows the share of completed interviews by the enumerator wherein the latter recorded the respondent as 'Not Working'.

Gender	Karnataka
Female(s)	Ms. X1 (0.8)
	Ms. X2 (0.7)
	Ms. X3 (0.7)
Males(s)	Mr. Y1 (0.7)
	Mr. Y2 (0.5)
	Mr. Y3 (0.5)

The flagged interviews against each enumerator are given below.

Surveyor Name	Start Time	End Time	District	Village	Household	Respondent Name
XXX	15 Feb. 11:08:36	17 Feb. 20:07:07	XXX	XXX	XXX	XXX
XXX	15 Feb. 13:51:44	17 Feb. 09:00:33	XXX	XXX	XXX	XXX
XXX	17 Feb. 14:32:27	20 Feb. 20:13:09	XXX	XXX	XXX	XXX
XXX	17 Feb. 8:36:45	17 Feb. 20:55:08	XXX	XXX	XXX	XXX
XXX	18 Feb. 10:36:36	18 Feb. 17:09:02	XXX	XXX	XXX	XXX
XXX	18 Feb. 11:13:10	18 Feb. 17:16:01	XXX	XXX	XXX	XXX
XXX	18 Feb. 14:26:01	18 Feb. 17:31:06	XXX	XXX	XXX	XXX
XXX	18 Feb. 10:21:43	18 Feb. 21:25:17	XXX	XXX	XXX	XXX
XXX	18 Feb. 11:16:04	18 Feb. 21:22:07	XXX	XXX	XXX	XXX
XXX	18 Feb. 11:34:21	18 Feb. 17:22:04	XXX	XXX	XXX	XXX
XXX	19 Feb. 11:10:18	19 Feb. 12:41:01	XXX	XXX	XXX	XXX
XXX	19 Feb. 8:55:54	20 Feb. 12:16:21	XXX	XXX	XXX	XXX
XXX	20 Feb. 10:12:45	20 Feb. 12:49:46	XXX	XXX	XXX	XXX
XXX	18 Feb. 13:05:12	18 Feb. 18:09:32	XXX	XXX	XXX	XXX
XXX	18 Feb. 11:36:50	18 Feb. 18:03:38	XXX	XXX	XXX	XXX
XXX	18 Feb. 12:06:18	18 Feb. 17:56:51	XXX	XXX	XXX	XXX
XXX	18 Feb. 9:53:16	18 Feb. 18:53:03	XXX	XXX	XXX	XXX
XXX	17 Feb. 12:13:53	18 Feb. 18:58:43	XXX	XXX	XXX	XXX