MIDS Working Paper No. 243

# How Good is Your Data?

## *Challenges of Ensuring Data Quality in a Large-Scale Survey: Lessons from the Tamil Nadu Household Panel Survey (TNHPS)*

**Kripa Ananthpur**
**Julie de Jong**
**Gargi Sridharan**
**Bhavya Shrivastava**

MIDS Working Paper No. 243
July 2023

# How Good is Your Data?

*Challenges of Ensuring Data Quality in a Large-Scale Survey: Lessons from the Tamil Nadu Household Panel Survey (TNHPS)*

**Kripa Ananthpur**
**Julie de Jong**
**Gargi Sridharan**
**Bhavya Shrivastava**

Madras Institute of Development Studies

MIDS Working Paper No. 243

How Good is Your Data?
*Challenges of Ensuring Data Quality in a Large-Scale Survey: Lessons from the Tamil Nadu Household Panel Survey (TNHPS)*

Kripa Ananthpur, Julie de Jong, Gargi Sridharan and Bhavya Shrivastava

Published by
**Madras Institute of Development Studies**
79, Second Main Road, Gandhi Nagar, Adyar
Chennai 600020 India
Phone: 2441 1574/2589/9771
pub@mids.ac.in | www.mids.ac.in

# How Good is Your Data?

*Challenges of Ensuring Data Quality in a Large-Scale Survey: Lessons from the Tamil Nadu Household Panel Survey (TNHPS)*

**Kripa Ananthpur[a], Julie de Jong[b]**
**Gargi Sridharan[c], and Bhavya Shrivastava[d]**

## Abstract

The growing popularity of population-based household surveys, especially in developing countries, brings with it its own set of challenges. These household surveys are expected to provide reliable estimates of the population that can influence policymaking. Thus, there is an increasing need to maintain the rigour of data collection by ensuring meticulous data quality practices so that the estimates are a true representation of the population. However, Indian statistical agencies, which have historically relied on paper-based interviewing techniques, are now faced with the challenges of adapting to computer-assisted interviewing (CAI) methods commonly implemented in the West. This paper highlights the experiences of transitioning from paper-based interviewing techniques to CAI methodology in the Tamil Nadu Household Panel Survey (TNHPS), a longitudinal survey of a large household sample of over 2 lakhs. In addition, this paper highlights the various challenges faced in ensuring data quality in large population-level surveys and provides quality control guidance to minimise error and increase data quality.

## Keywords

TNHPS, Panel survey, Data quality, Quality control, Large sample, CAI.

[a] Associate Professor, MIDS (Corresponding Author). Email: kripa@mids.ac.in
[b] SMRFS Deputy Director for Survey Operations, ICF
[c] Research Scholar, MIDS
[d] Research Scholar, Trinity College Dublin

## Introduction

Population-based household-level surveys are an important contribution to social science research in the twenty-first century. These surveys provide valid and reliable population-level measures and are a valuable source to inform policy decisions. However, in recent years the quality of data from household surveys has been declining because of both measurement errors and nonresponse bias (Meyer et al., 2015). This issue is especially germane in India, where several large surveys have come under scrutiny. For example, the National Sample Survey Office (NSSO)'s Consumption Expenditure Survey conducted in 2017–2018 was not released, allegedly due to its data quality concerns (Seshadri, 2019). Furthermore, the Indian official statistic systems suffer from the challenges of adapting to ICT (information and communication technology) based solutions for data collection that are comparable to global standards (Kumar et al., 2020). The significant transition from paper-based surveys to computer-assisted interviewing (CAI) provides an opportunity to address a series of challenges (Sastry et al., 2021). CAI refers to a set of data collection techniques that include both face-to-face digital interviewing called computer-assisted personal interview (CAPI) and remote interviewing through web-based applications called computer-assisted telephone interview (CATI). As governments and data collection agencies in India are moving towards CAI at the moment, its implementation is valuable as it can significantly improve the speed of data entry and delivery, identify and resolve data entry errors and inconsistent reports by respondents, and capture additional information about the interview, including audio files and geographical coordinates. As the complexity and length of the questionnaires increase, so do the benefits. However, while CAI provides enormous opportunities to improve the quality of data in the long term, its adoption poses some challenges during the transitional period.

The Tamil Nadu Household Panel Survey (TNHPS)[1] is a state-wide panel survey that aims to measure changes in income, occupation and

[1] The TNHPS–PBS was a joint collaboration between the MIDS, the Department of Economics and Statistics (DES), the Government of Tamil Nadu (GoTN) and the Survey Research Center (SRC) at the University of Michigan.

other outcomes at the household level through multiple waves of a longitudinal panel survey in Tamil Nadu. The first phase of the TNHPS was expected to conduct a Pre-Baseline Survey (PBS) in 2018–2019 with a sample of approximately 2.12 lakh households and around seven lakh individuals across the state. The goal of the TNHPS–PBS was to collect basic demographic and socioeconomic information on each household in the sample both to provide state- and district-level estimates of key indicators and to serve as the sampling frame for the first wave of the panel survey, the Baseline Survey (2022–2023), which is currently ongoing. However, the pandemic delayed the commencement of the Baseline Survey (2022–2023) because of the lockdown and obvious health concerns. This provided the opportunity for a spin-off study on a subset sample from the large PBS database to study the socioeconomic impacts of the pandemic (Ananthpur et al., 2022). The study, called the Tamil Nadu Covid Pulse Survey (TNCPS), had a longitudinal study design with a sample size of around 13,000 households in 2020 and 2021. Being one of the few CATI surveys during the pandemic, the TNCPS also adopted a series of quality control measures. Because of the presence of the large sample size of the PBS and the TNCPS, and the fact that these surveys were the first large-scale data collection initiative of Government of Tamil Nadu (GoTN) using the CAI methodology, integrating stringent quality control measures was critical to ensuring high-quality data. Using the experience of the TNHPS, this paper highlights the various challenges faced in ensuring data quality in large population-level surveys and provides quality control guidance to minimise error and increase data quality.

We first provide an overview of survey quality and related frameworks, using specific case studies to illustrate the operationalisation of these in real-world surveys, followed by a description of our approach to the application of quality frameworks to the TNHPS–PBS. In the subsequent sections, we provide a detailed description of the quality control practices followed in the TNHPS–PBS, the challenges faced and finally lessons for future surveys in countries with nascent survey research infrastructure.

## Survey Quality and Quality Frameworks

Central to the survey lifecycle is the concept of *survey quality*. While early researchers recognised the vulnerabilities of survey data to errors primarily associated with sampling, conceptualisations of error were expanded in the mid-twentieth century by Deming (1944), whose typology included thirteen factors that he believed were critical to assessing survey quality (see also Groves & Lyberg (2010) and Lyberg & Stukel (2017)). Subsequently, many frameworks and models have emerged to specify and estimate both statistical and non-statistical indicators of data quality. Nowadays, several paradigms are generally used in assessments of survey quality, including (a) total survey error (TSE), (b) fitness for the intended use and (c) monitoring survey production process quality, all of which can be affected by survey infrastructure, costs, respondent and interviewer burden, and other study design specifications.

### *Total Survey Error*

Total survey error (TSE) is widely accepted as an organising framework in the design and evaluation of surveys (Groves et al., 2009). Errors in survey estimates consist of variances of estimates (reflecting estimate instability over conceptual replications) and systematic deviations from a target value ('biases') and can be errors of both observation and non-observation. TSE defines quality as the estimation of and reduction in the mean square error (MSE) of the statistic(s) of interest – the sum of random errors (variance) and squared systematic errors (bias) – although the MSE for each statistic in a survey is not typically estimated (Vehovar et al., 2012). TSE takes into consideration both measurement (construct validity, measurement error and processing error), that is, how well survey questions measure the constructs of interest, and representation (coverage error, sampling error, nonresponse error and adjustment error), that is, whether one can generalise to the population of interest using sample survey data (Groves et al., 2009). Errors associated with measurement are generally observable and involve differences between reported values of a survey variable and some underlying or generally accepted 'true' values. Potential sources of such errors include interviewers, respondents, the mode of data collection,

aspects of the questionnaire and/or associated translations as relevant, and data processing (Groves & Lyberg, 2010). In contrast, errors of representation are non-observable and result from inconsistency between the population of interest and the population from which the data were collected. From the TSE perspective, there are cost-error trade-offs, that is, there is tension between reducing these errors and the cost of doing so.

### *Fitness for Intended Use*

Although widely used, the TSE framework has been criticised for its lack of user perspective. Therefore, the 'fitness for intended use' paradigm can be used to supplement the TSE framework (Biemer & Lyberg, 2003). The fitness for intended use is a multidimensional approach that focuses on the quality assessment criteria in terms of the degree to which survey data meet user requirements. In this approach, the study design strives to meet the needs of users in terms of survey data accuracy and other quality dimensions including comparability, relevance, accuracy, timeliness and punctuality, accessibility, interpretability and coherence. In this perspective, ensuring quality on one dimension (e.g. accuracy) may conflict with ensuring quality on another dimension (e.g. timeliness); and there may be difficulty in meeting user needs in terms of both TSE and fitness for intended use. However, the overall aim is to optimise quality, minimise costs and burdens, and recognise and document design constraints at all levels.

### *Monitoring Survey Production Process Quality*

Monitoring survey production process quality is an iterative process that emphasises the concept of continuous process improvement (Groves et al., 2009), focusing on quality at three levels: the product, the process and the organisation (Lyberg & Biemer, 2008). Product quality is the expected quality of survey deliverables, which is often determined by data users. Process quality refers to the quality of the processes generating the product, through selection, measurement and analysis of process variables relevant to the particular survey (Lyberg & Stukel, 2010). Ensuring production process quality requires the use of quality

standards and the collection of standardised study metadata, question metadata and process paradata (Couper, 1998) and is operationalised through the quality control process guided by quality planning and assurance. Paradata include specific survey characteristics such as the start time, the end time, interviewer characteristics and so on that enhance quality control processes, especially in large-scale surveys (Goel et al., 2022). The quality control outcome measures result in a quality profile that can also be used to make recommendations for improvements and is subsequently reflected in future planning. Organisational quality refers to the features that make good processes possible, such as quality-oriented top management, good user relationships, constancy of purpose and competence development programmes.

Each of these approaches to assessing survey quality has strengths and weaknesses. TSE alone is insufficient for survey quality because of its lack of user perspective. As Biemer and Lyberg (2003) argue, the TSE framework can be supplemented with the 'fitness for intended use' framework, which lends an aspect of practicality and provides a general framework for assessing quality, integrating TSE with the accuracy dimension, while the survey process quality acknowledges the critical effect of processes on the actual data. These three paradigms can be integrated into a comprehensive quality assurance and quality control framework to assess survey data quality through the development of protocols and procedures to minimise the errors from different sources that may affect data quality (Hansen et al., 2016). For example, the World Health Survey employs a total quality management paradigm that examines the survey process at each step, including survey instrument design, sampling, translation, training, survey implementation, data capturing, data analysis and selection of quality indicators (Üstun et al., 2005). Integrative approaches are also used in many other surveys, during both the data collection period and post-field operations through internal quality audits and external quality assessments (Börsch-Supan et al., 2008; Vila, Cervera, & Carausu, 2013; Wuyts & Loosveldt, 2019).

There are many potential sources of error in the design and execution of survey operations. In interview-administered surveys, interviewers have the potential to both contribute to and minimise nonresponse bias and measurement error while playing a key role in

sampling frame construction through a selection of sampling units, gaining respondent cooperation and following protocols during the questionnaire administration process. Therefore, a significant number of quality assurance and quality control processes are focused on interviewer training and monitoring. In the following, we consider several case studies using a variation of an integrated strategy to monitor interviewer performance and then discuss how we modified this approach to address the challenges of ensuring data quality during digital data collection in the TNHPS–PBS.

## *Indian Case Studies*

Rigorous survey research employs quality control measures informed by one or more quality frameworks during the data collection period to guide interventions aimed at reducing interviewer-induced errors (Mneimneh et al., 2019). Survey organisations generally use multiple approaches to quality control, including both interview verification and identification of errors in specific cases (e.g. flagging interviews of unusually short duration).

Indian panel surveys such as the Consumer Pyramids Survey[2] (CPS) and the Television Audience Measurement (TAM)[3] Panel use many monitoring processes. Launched in 2008, the CPS is a longitudinal survey of 1,70,000 Indian households, designed to measure changes in the economic well-being of households over time (Mneimneh et al., 2019). Several characteristics of the survey design called for comprehensive data quality measures, including the extensive geographical spread of the sample, adverse climatic conditions and poor transportation and infrastructure, all factors that increase the likelihood of interviewer deviation from study protocol and data quality vulnerability. To combat the threat of poor-quality data, the data collection organisation, the Centre for Monitoring Indian Economy (CMIE), deployed three quality control procedures: (a) verification, (b) data-driven assessment and

[2] The Consumer Pyramids Survey collects continuous data on households on their occupation, income, consumption, aspirations and other socioeconomic indicators.

[3] The Television Audience Measurement Panel, which was established in 1998, used model-based quality control methods.

(c) respondent mailings. The first procedure involved interview verification through both back-check telephone calls and back-check in-person visits by the field investigating officer (FIO). The FIO verified that the interviewer did indeed conduct the survey, confirmed the duration of the survey and collected information on the interviewer's professional behaviour. The data-driven assessment involved a validation team, which functioned from the headquarters to run high-frequency checks and validate all the data collected by the FIOs and the interviewers in real time as the data were uploaded to the system. This assessment monitored several types of data for all the households, including GPS coordinates to verify the movement of the interviewer in the field and keystroke and time stamp data assessment of navigation through the instrument as indications of the veracity (or potential fabrication) of the survey data. Lastly, the respondent mailing procedure entailed sending out 'thank you' letters to the households that participated in the survey. These letters were delivered to the addresses and acknowledgements were received. Undelivered letters returned to the sender were investigated to determine whether causes were associated with benign issues such as residential relocation or were evidence of malpractice in the survey execution (Mneimneh et al., 2019).

The TAM panel in India, which was established even before the CPS in 1998, used quality control measures similar to those used in the CPS. More recently, the TAM has also relied on model-based quality control methods. The TAM passively collected data through a meter fixed to respondent household televisions. Because the data were passively collected, the interviewers played a unique role by motivating the respondents to participate and adhere to the study protocols. These interviewers also helped in recruiting a refresher sample when required by the study. They were also responsible for collecting and updating household-level information at the end of each calendar year. To supervise the teams of interviewers, TAM installed a regional field head (RFH) in each state of India who was responsible for field monitoring of the interviewers, providing refresher training to the interviewers and providing performance feedback after conducting back checks. Model-based quality control processes were implemented by auditing

a small sample of households with the help of an external consulting firm under the direction of the 'Measurement Science' department. This small sample for model-based quality control consisted of households that either were selected randomly or showed data inconsistency. Both substantive data and paradata were subjected to data-driven assessments conducted weekly to detect unusual patterns, and multilevel statistical models were incorporated into interviewer monitoring processes to facilitate the investigation of interviewer falsification. Traditional methods of field verification in tandem with data-driven approaches were effective in increasing efficiency and data quality and decreasing costs through the detection of deviations from protocols and subsequent intervention (Mneimneh et al., 2019).

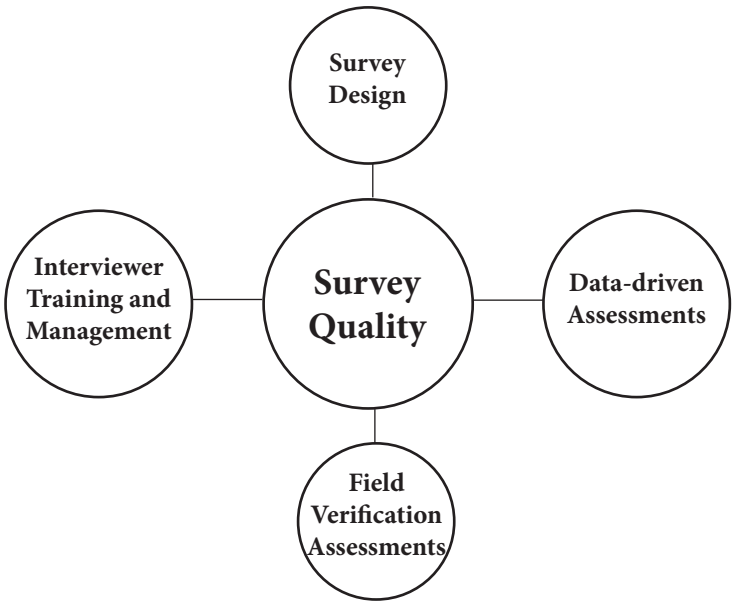
In recent years, the role of data-driven approaches of using paradata (the usage of dashboards to monitor progress, flag issues and so on) to improve survey quality has been tested in more Indian surveys such as the India Working Survey (IWS). However, the IWS dealt with a small household sample of approximately 4000 households. Goel et al. (2022) highlight the need for more important forms of paradata such as GPS coordinates that are often more accurate in identifying any non-standard behaviour in the field.

While the aforementioned approaches have been discussed mostly for in-person surveys or CAPI, a great deal of innovation in these approaches became essential during the Covid-19 pandemic. This period saw a surge in remote interviews such as telephonic surveys or CATI as mentioned previously. This transition, although commendable and necessary, poses serious threats to the quality of data collected. In the absence of rapport building with the respondents, in-person verifications of data and so on, the need for strengthening data quality using control measures is more significant in a CATI model. Having said that, the TNHPS, through its CAPI and CATI approaches, addresses some of the underlying challenges by enforcing paradata-related quality control measures in the context of a developing country. Further, the large sample size also provides a framework to adopt a combination of survey design, interviewer training, and paradata-driven assessments in addition to field monitoring to achieve optimum data quality.

## TNHPS–PBS Quality Control Framework

In the TNHPS–PBS, we adopted a similar integrated approach for quality assessment as illustrated in Figure 1, which details how the components of quality control are informed by the theoretical approaches to quality.

**Figure 1. TNHPS–PBS Quality Framework**




### *TNHPS Pre-Baseline Survey (PBS) Design*

Collection of high-quality data begins with a rigorous study design and management. The TNHPS–PBS covered all 32 districts[4] in the state, selecting a total of 573 Census Revenue Villages and 678 Urban Frame

[4] When TNHPS was initiated in 2017, there were 32 districts and the samples were drawn from these 32 districts. Subsequently, between 2018 and 2020, the state was reorganised and currently there are 38 districts in the state. Tirunelveli, Vellore, Kancheepuram, Villupuram and Nagapattinam districts have been either bifurcated or trifurcated to create six new districts. Thus, we are considering the erstwhile districts in our analysis.

Survey Blocks (UFSBs) as primary sampling units (PSUs).[5] Each PSU was listed and the interviews were completed in households that provided consent. In total, interviewers contacted more than 2,40,583 households and completed interviews in approximately 92% of occupied residential dwellings, for a total of 2,12,282 completed household interviews, which included information on 7,45,653 individuals living in these households (Ananthpur et al., 2021; Sastry et al., 2021).

The TNHPS–PBS marked a major transition of the state's data collection processes and procedures from paper-assisted personal interviewing (PAPI) to CAI, with interviewers using electronic tablets installed with SurveyCTO, a mobile data collection platform that can be used both online and offline with multiple layers of in-built encryption. SurveyCTO provides functionality for monitoring the incoming data daily and a visualisation platform to facilitate data monitoring and analysis. Interviewers administered a comprehensive questionnaire (around 30 minutes) to the key informant in each household, collecting data both at the household level (income, asset/land ownership, availability of basic infrastructure amenities) and at the individual level (demographic data), for all residents of the household. Upon completion of the interviews, these data were transmitted to the server using the internet. These data were stored securely on the server and were only accessible to the study team using confidential login details.

## *Interviewer Training and Management*

The PBS implemented a face-to-face mode of data collection, rendering interviewers an indispensable part of the data collection process but also requiring close monitoring and deliberate coordination between the interviewers and researchers to ensure high quality at every level of data collection. Considering that this was the first CAI survey undertaken by the staff, a 'training of trainers' (ToT) detailing all the survey proceedings was first provided to the supervisory staff who were primarily involved in the overall administration and regulation of the survey. Following this,

[5] The Census Revenue villages are the villages listed in the Census of India 2011. The UFSBs are the primary units used in the Urban Frame Survey 2015 conducted by the National Sample Survey Office (NSSO). On average, PSUs in both rural and urban areas were comprised of about 180 households.

the field manager and the training team conducted subsequent rigorous training for the interviewers and supervisors in each district before the initiation of the TNHPS–PBS. The primary focus of the training sessions was to equip the interviewers with the skills to undertake CAI and related software and hardware management. The training emphasised the importance of error-free data collection, standardised interview protocols, conceptual and definitional clarifications and other monitoring mechanisms to be followed. Considering the longitudinal nature of the TNHPS, the training sessions also focussed on the process of collecting accurate contact details of respondents and friends/relatives for future tracking and GPS coordinates of dwellings to aid in the identification of the households during subsequent waves, thereby minimising possible attrition in data collection.

The PBS deployed about 680 interviewers across 32 districts. As stated previously, the interviewers were monitored by the supervisory staff who worked with the researchers to oversee the data collection progress and the performance of interviewers across all districts. This comprised 350 primary supervisors who monitored the daily productivity and efficiency of the interviewers in districts. In addition, considering the large sample size, each district was allocated a district-level field manager (DFM) to facilitate the survey operations. Additionally, a regional-level field manager (RFM) was also assigned to each of the six regions[6] of the state to monitor the survey progress and performance of the districts in the respective region. Thus, the secondary supervisory team consisted of 6 RFMs and 32 DFMs who worked in tandem with the primary supervisors who were hired separately and responsible for addressing the day-to-day concerns of the interviewers, providing regular feedback on the interviewer's performance and working in tandem with the researchers in conducting quality checks and providing refresher training to interviewers as and when required.

[6] The six regions of Tamil Nadu are Chennai, Coimbatore, Cuddalore, Madurai, Salem and Tirunelveli.

In rural areas, interviewers canvassed the entire sampled village to note approximate population concentrations, referring to pre-existing village maps available from each Village Administrative Officer or integrated child development centre. In urban areas, interviewers replicated the process for each of the selected UFSBs, using the UFSB maps designed by the NSSO for each urban block. They marked the boundaries and area names, noted permanent landmarks that served as boundaries along with other identifiers on these maps and uploaded images of the same. The next step in the data collection process was to determine whether each PSU would be enumerated in its entirety or not. In each selected PSU with a population estimated to be less than 1200 (based on observation and available maps), all households were contacted with a request to complete the PBS. In those PSUs where the estimated population was 1200 or more, the interviewers divided the village into 'hamlet-groups' or 'sub-blocks' and selected two hamlets—the largest hamlet (hamlet group 1 [HG1]) by default and one other hamlet which was randomly selected (hamlet group 2 [HG2]) for complete enumeration.

This process, called the 'Hamlet/Sub-block Creation', was modelled on the standard methodology followed by the NSSO. However, the PBS was the first effort to deploy this process in the field through a CAI framework. This entire process was digitised for the interviewer to identify the PSU, estimate the population size with the help of a Village Administrative Officer (VAO) in the case of villages and other knowledgeable persons in the case of UFSBs, trace the boundaries of the PSU using the maps and code them by serial numbers. Using this information, the CAI software identified the largest HG1 by default and selected the HG2 at random. Further, the maps were also documented by images taken by the interviewers in the field. This process facilitated a more accurate and systematic collection and verification of these data.

### *Interviewer Monitoring*

Interviewers were monitored through a series of both field verification and data-driven assessments.

## Field Verification Assessments

### *Spot Checks*

Spot checks were implemented as a monitoring mechanism to evaluate an interviewer's performance on various parameters. Each spot check was conducted simultaneously by two supervisors at the same time, using spot check forms that were developed for electronic capture. Similar to the survey forms, these forms were also deployed on the tablets using SurveyCTO and uploaded to the servers. During the fieldwork period, every interviewer was assessed at least once using the spot check form either while conducting the survey or while completing a hamlet[7] creation form.

The supervisory staff scored the interviewer on a scale of 1 (poor) to 5 (excellent) based on his/her ability to: (a) identify the key informant (an adult member of the household); (b) convince the respondent to participate in the survey; (c) address respondent concerns; (d) administer both informed consent for the survey and separate consent for audio recording; (e) adhere to the interview protocols throughout the survey; (f) accurately capture the GPS coordinates; and (g) physically label the door/adjacent wall of the household with the TNHPS number and appropriate result codes[8]. The field staff also evaluated the interviewers on a subjective basis, scoring interviewers on whether questions were asked verbatim (as required by protocol), the interviewer's ability to adapt to CAI, handling of the tablet, and efficiency and the number of interviews completed per day and so on.

Both types of spot check data were assessed in near-real time, identifying interviewers who consistently committed more errors. The researchers transmitted this information to the field team and those interviewers with poor performance received feedback and refresher training and were subsequently subjected to more frequent spot checks.

---

[7] Interviewers were responsible for assessing whether the PSU was large enough that it required subdivision into hamlets in rural areas or into sub-blocks in urban areas.

[8] The TNHPS research team developed result codes to identify the various possible outcomes for each contact attempt made by the interviewers while conducting the PBS.

As the supervisory staff scored these interviewers, the average of the two scores was calculated to assign a final score to each interviewer.

**Table 1. Spot Check Scores**

| Score | Prevalence rates |
|---|---|
| 1 | 2.0 |
| 2 | 22.1 |
| 3 | 55.4 |
| 4 | 19.7 |
| 5 | 0.9 |

Source: Estimates from the Spot Check Analysis

Around 76% scored 3 or higher score, indicating that the majority of the interviewers adhered to the protocols. Of the 923 interviewers deployed for data collection across the 32 districts, around 55% received a score of 3. The percentage of interviewers with a low score of 2 and below was about 24%. The spot check analysis showed that the interviewers generally seemed to fare well both in conducting the main survey and in creating the hamlet form from the perspective of the supervisor who was observing their steps in the field. It is important to note that the secondary supervisory team of DFMs and RFMs was independent of the primary data collection unit, which consisted of the interviewers and primary supervisors to ensure bias-free supervision of the data collection process. The average score obtained by the interviewers based on their performance of creating hamlet groups was 3, 'Good', while the score for conducting the main TNHPS–PBS was nearly 3 (2.98). However, it is to be noted that the goal of the spot check was to ensure that the interviewers followed the aforementioned protocols to maintain standardisation across all sample respondents in the districts. It is important to note that the spot-check data are limited by the subjective nature of the spot check measures, and thus a series of objective processes were also implemented, as we will discuss in the next sections.

### *Back Checks*

One objective measure that we implemented was back checks, which played an important role in the verification of survey quality and validity. The back check process was conducted using an electronic instrument administered through SurveyCTO during an in-person visit with only a limited number of assessment questions. As households were assigned a final result code and associated data were uploaded to the SurveyCTO server, a small sub-sample (approximately 10% of the total surveys completed) was selected at random by the researchers and was back-checked by the supervisors through the use of an electronic form. Staff first confirmed that the house had been contacted for a survey, and further verified: (a) the number of household members residing within the household, (b) the social group of the household head, (c) ownership of agricultural land and (d) the presence of a household member with a disability. Finally, the staff member enquired whether the interviewer had completed the interview using a tablet, a paper-based instrument or both, as there were reports of some interviewers who were reluctant to use tablets and instead used paper-based instruments. Unlike the traditional back check methods, wherein the supervisor checks the data collected by the interviewer during their field visit, the digital back check process followed in the PBS was tweaked to reduce bias. The supervisor did not have access to the data collected by the interviewer and collected data afresh for a certain number of questions during back checks. The variations in the data collected by the interviewer and the supervisor could only be noticed at the back end by the research team. This process was followed for both primary- and secondary-level supervisors.

These data were used to verify adherence to the data collection protocol and to detect any interviewer-attributable errors in the recording of respondent responses to survey items. It was important to have the supervisors do the back checks independently, without access to the original survey data, to reduce bias.

The research team conducted a total of 11,600 back checks at the state level, with an average of 16 back checks for each interviewer. Errors were identified in about 12% of completed surveys. The error percentage across the districts was variable, ranging from 4% to 26%. About ten

districts had an error rate of less than 10%, while the remainder had error percentages above 10%. The interviewers were required to re-interview those households where the back checks uncovered one or more errors. The TNCPS, which was a CATI-based survey during the pandemic, also implemented a telephonic framework of the back check design as part of its quality control.

### *Audio Verification*

While the conventional methods of ensuring data quality such as spot checks and back checks offer insights into the interviewer's performance on pre-specified parameters, sophisticated tools such as computer audio-recorded interviewing (CARI) are crucial for studying the behaviours of both the interviewer and the respondent, assessing the survey atmosphere, the delivery of questions and so on. Audio files produce a wealth of qualitative data on the interaction between the interviewer and the respondent, which are often used for the triangulation of data checks. Additionally, this technology helps field managers and researchers detect any falsification of interviews (Thissen et al., 2007), obtain useful data for intervention and conduct further training to ensure minimised interviewer errors in the survey. The TNHPS–PBS adopted an extensive CARI methodology in its processes wherein consent was sought specifically for the audio recording of interviews from the key informants, which in addition to the use of CAI was a new methodology for the interviewers. Refusal of consent for audio and video recording of interviews was shown to be high in the south Indian context (Chauhan et al., 2015), and similarly, in the TNHPS–PBS, only 26% of key informants, that is, approximately 54,000 households, agreed to the audio recording of interviews. Privacy issues and related fears were seen as the primary concern for key informants. In interviews where consent was obtained, several sections of the survey were audio-recorded and the digital audio files were electronically stored. The data collection software picked a section at random to record, thus preventing the interviewers from manipulating the system by modifying their behaviour only during that specific section. Interviewers were unaware of when they were recorded. Audio monitoring helped the PBS team

identify positive behaviours such as consistency, transparency and adherence to protocol among the interviewers, as well as concerning issues such as unnecessary paraphrasing, lack of probing and so on, all of which facilitated further interviewer training.

Researchers were responsible for transcribing the audio files and auditing them. The audits were quantified and the interviewers were scored on a numeric scale from 0 to 20, based on the following parameters: whether (a) the key informant was accurately identified by the interviewer; (b) the consent script was duly read to the key informant (automatically recorded for all interviews); and (c) the survey questions were administered as per the protocols established. These protocols included reading the questions verbatim, providing the standardised definitions to the respondents, explaining the option categories without skipping and so on. Audio files were assessed at both the district level and the interviewer level, and for each interviewer in each district, a sample of five households were selected taking into account variables such as start time, duration of the interview, and area type and audio consent. This resulted in an average of four audio audits per interviewer, resulting in audits for about 1% of all households where an interview was attempted (inclusive of the households that may have denied consent).

As stated previously, the results from the audio audits were also used to supplement other data checks. For example, in cases where interviews were flagged for being relatively short (less than five minutes) in the data-driven assessment discussed below, audio audits were useful in identifying the reasons for the short duration. For example, the audio files from such entries were analysed to identify instances of rushed questioning and other ambiguities. For those interviewers with a large number of audio consent refusals, this method helped in understanding the reasons behind them.

Similar to the pattern borne out in the other field assessment methods, audio verification revealed significant variation across the districts, which was mainly due to low consent rates, wherein only five districts (out of 32) had the highest number of audio audits conducted. In general, there was a high consent refusal by the respondents for audio recording of the interviews which is to be expected. In spite of the low consent rate, the results from the audio checks were useful in

supplementing the findings from the other checks that were undertaken. These results were shared periodically with the concerned personnel to facilitate feedback and intervention for interviewers.

The low audio consent rates implied that a rigorous comparison across the districts or interviewers was not possible because of potential issues of bias. While in other contexts a high number of consent refusals itself could be viewed as a proxy measure of the interviewer's performance, such assumptions may not hold true in this case as there is an inherent aversion among respondents to being recorded, which might have led to high number of consent refusals. Hence, granular assessments based on audio audits were not possible in this case. This is also reflective of the drawback of audio audits as a quality control tool in developing countries where cultural challenges can limit capturing of audio recordings in large-scale digital surveys. However, the overall audio recordings from the districts and interviewers facilitated efficient identification of recurring issues that would otherwise have been missed, with a significant impact on the data quality. The TNCPS also had a CARI design; however, employing the same in a CATI framework was more challenging than in a CAPI framework. The quality of the audio recorded depended on the quality of the cellular connection, the phone or the landline of both the interviewer and the respondent.

In addition to the survey data, CARI was also used to ensure that consent was duly obtained by the interviewer. Obtaining consent plays a pivotal role in social surveys considering the plethora of sensitive information collected from the survey respondents. To ensure this, the section consisting of the consent script following the question was audio-audited. A random subset was checked for each interviewer to verify this process.

### *Data-driven Assessments*

In addition to adopting approaches to quality control through field verification methods, the researchers also carried out several data-driven assessments to enable the effective detection of data outliers during the process of data collection itself, a distinct advantage of the digital data collection mode. Outliers were further investigated and

remedied, as warranted. This was carried out in two ways. The pre-existing dashboards on the SurveyCTO Server Console were used by the researchers in the initial phase of the survey to draw cross-tabulations to test for missing values, inconsistencies and so on. However, this process could only be carried out in the initial phase with a limited sample size as the server failed to process these analyses as and when more observations were completed and sent to the server. This happens to be a limitation of the Survey CTO Console, which may be more amenable for small samples than for large-scale surveys. More statistical analyses were conducted by the researchers using software such as STATA on a biweekly basis to identify various inconsistencies in the data. Less than 2% of the sample had such inconsistencies. For example, there were some cases wherein the household head was not accurately identified in the initial survey, leading to erroneous information on the social group, religion and migration status, variables for which data pertaining only to the household head were collected. In other cases, more than one household head was indicated in a particular household, again leading to inconsistencies in the dataset[9]. Additionally, as stated previously, the duration variable for each interview and sub-section was also used to flag interviews that were completed well within the duration of an average interview. Using several pre-tests and pilot studies, the average time of an interview was determined to be 12 to 15 minutes. And this duration per interview usually decreases as the interviewer has more practice with data collection (Olson & Smyth, 2020). Thus, with the help of timestamps, interviews completed within 10 minutes were examined for other flags. If the data were found inadequate as identified through data-driven assessments, the interviews were re-administered by the interviewer.

## *Nature of Interventions*

The PBS followed various interventions by the supervisors to the interviewers to address data quality issues. The high-frequency checks,

[9] Some of the issues encountered in the data-driven assessments were a product of the limitations of the data collection software at the time of the TNHPS–PBS. See Sastry et al. (2021) for a detailed discussion of considerations when selecting data collection platforms.

as stated earlier, flagged minor issues concerning wrong entries of sample IDs, enumerator IDs, etc. that are common in data collection. These issues were simply communicated to the interviewer and rectified at the back end. In cases of serious data issues flagged by the high-frequency checks, such as inconsistencies, errors in income data, asset data, and so on, the research team shared the data with the supervisor, who verified the errors and conveyed them to the interviewer. If these issues persisted, the interviewer was temporarily withdrawn from data collection and was given face-to-face retraining by the supervisor in the district. This also applied to the interviewers whose data were found to be inaccurate during back-checks. As opposed to the high-frequency checks and the back checks, the audio checks flagged more serious data and protocol breaches that needed more critical interventions. Periodic meetings with the higher-level personnel in charge of data collection were arranged to facilitate listening to the audio files from the field. Audio files that reflected both excellent and poor interviewer performance were presented to the higher-level personnel, the primary and secondary supervisors. Interviewers breaching standard survey protocols such as interviewing minors, using paper questionnaires, and so on were reprimanded by the higher management and sometimes even removed from data collection.

In collaborative surveys such as the PBS, the process of providing feedback is delicate because of the presence of various tiers of individuals working towards a common goal. But given the necessity of periodic intervention in surveys, feedback must be delivered in a way that promotes interviewer morale, often emphasising better data quality.

## Conclusion

In this paper, we have addressed the challenges of ensuring high-quality data, especially in large-scale surveys, and have identified the numerous ways in which the usage of CAI resulted in improvements to the overall data quality. In contexts where the issue of data quality has become contentious, large-scale surveys must adopt multiple quality control measures that complement each other to minimise data quality concerns. This is particularly important in a context such as India, where

population-based surveys are frequently large in scope and require a significant number of interviewers and associated oversight. The shift from PAPI to CAI, while initially challenging, also provides additional, tighter layers of quality control. This ranges from the rapid uploading of survey data that allows researchers to assess quality in real time and taking immediate measures as opposed to PAPI, which requires data entry to be completed to access and assess the data quality which prevents simultaneous, real time rectification of data collection errors; audio audits that greatly enhance data quality and spot interviewer and respondent bias; GPS coordinates that validate the enumerator's visit to the field and also make future visits easy and so on. In addition, advances in CAI software promise continued opportunities for data quality mitigation efforts (Sastry et al., 2021).

However, the adoption of CAI is not without challenges, chiefly among them ensuring that the data collection software provides optimal functionality for the specific project to increase the likelihood of study success and the collection of high-quality data. For example, at the time of TNHPS–PBS data collection, SurveyCTO did not have an independent sample management system, which resulted in difficulties in timely assessments of interviewer contact attempts and subsequent monitoring of nonresponse and associated bias. Other software constraints experienced while implementing the PBS, including limitations of case management, quality assurance and control functionalities, also illustrate the importance of selecting data collection software that aptly fits the needs of the project. Another challenge was the availability of internet in a few remote spots in Tamil Nadu, especially in the Nilgiris district. A lack of connectivity meant that the interviewer could not send data readily to the server after collection and had to wait to return to the block office, which had Wi-Fi connectivity. This, however, was a minor concern restricted to some remote rural areas. A successful transition to a CAI system also requires highly skilled personnel who are trained in data security considerations, technicalities, data collection processes and management expertise. Further, attitudinal changes and monetary investments in human resources play a vital role in facilitating this transition in developing countries. As seen in the TNHPS, the gains in

data quality realised through the successful adoption of CAI have the potential to be significant for governments and other data agencies.

In the context of data-based governance, ensuring high-quality data becomes even more important as it has direct implications for governance. With more states moving towards evidence-based policymaking and shifting towards digital data collection methods, the experience of TNHPS in ensuring high-quality control of digital data becomes invaluable. The TNHPS–PBS provides a useful example of an integrated approach to quality control and assessment in a large-scale survey and can serve as a useful guide for other large-scale digital surveys.


## *Acknowledgements*

We extend our gratitude to the GoTN for the opportunity to conduct the TNHPS. We also thank our partner, the Department of Economics and Statistics (DES). We thank the anonymous referee for the comments and our colleagues at the Madras Institute of Development Studies (MIDS) and the Survey Research Center (SRC) for their contributions. The authors take responsibility for any errors that may remain.


## *Conflict of Interest*

The authors declared no potential conflicts of interest. This paper is a part of the TNHPS–PBS research funded by the GoTN between 2016 and 2019 and undertaken by the MIDS in association with the SRC and the DES. The views expressed in this paper are those of the authors alone and do not reflect the opinions of the GoTN. The corresponding author is one of the principal investigators of this research.

## References


Ananthpur, K., Jafar, K., Sridharan, G., & Venkatachalam, L. (2022). *Impact of the Covid-19 waves on Tamil Nadu's economy and society* (MIDS working paper no. 241). Madras Institute of Development Studies. https://www.mids.ac.in/assets/doc/WP_241.pdf

Ananthpur, K., Venkatachalam, L., Jafar, K., Axinn, W. G., & Sastry, N. (2021). *Tamil Nadu Household Panel Survey, Pre-Baseline Survey: State Report*. Madras Institute of Development Studies, Government of Tamil Nadu and Survey Research Center, University of Michigan.

Biemer, P., & Lyberg, L. (2003). *Introduction to survey quality*. John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/0471458740

Börsch-Supan, A., Brugiavini, A., Jürges, H., Kapteyn, A., Mackenbach, J., Siegrist, J., & Weber, G. (2008). *First results from the survey of health, ageing and retirement in Europe (2004–2007)*. Mannheim Research Institute for the Economics of Aging (MEA). https://iris.unive.it/bitstream/10278/22179/3/Brugiavini_Pasini_Peracchi_FRBnew.pdf

Chauhan, R. C., Purty, A. J., & Singh, N. (2015). Consent for audio-video recording of informed consent process in rural South India. *Perspectives in Clinical Research*, *6*(3), 159–162. https://pubmed.ncbi.nlm.nih.gov/26229752/

Couper, M. (1998). Measuring survey quality in a CASIC environment. *Proceedings of the Survey Research Methods Section*, 41–49. https://cir.nii.ac.jp/crid/1572824500641137920

Deming, W. E. (1944). On errors in surveys. *American Sociological Review*, *9*(4), 359–369. https://www.jstor.org/stable/2085979

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, *74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Goel, D., Abraham, R., & Lahoti, R. (2022). *Improving survey quality using paradata: Lessons from the India working survey* (IZA Discussion Paper no. 15041). IZA Institute of Labour Economics. https://repec.iza.org/dp15041.pdf

Hansen, S. E., Benson, G., Bowers, A., Pennell, B.-E, Lin, Y., Duffey, B., Hu, M., & Hibben, K. C. (2016). *Survey quality: Guidelines for best practice in crosscultural surveys*. Survey Research Center, Institute for Social Research, University of Michigan. https://ccsg.isr.umich.edu/chapters/survey-quality/

Japec, L. (2006). Quality issues in interview surveys: Some contributions. *Bulletin of Sociological Methodology / Bulletin de Méthodologie Sociologique*, *90*(1), 26–42. https://doi.org/10.1177/075910630609000104

Kumar, A., Mohanan, P. C., & Kundu, A. (2020). Disastrous decade for data: The way forward. *Economic and Political Weekly*, *55*(41), 18–21. https://www.epw.in/journal/2020/41/commentary/disastrous-decade-data.html

Lyberg, L. E., & Stukel, D. M. (2017). The roots and evolution of the total survey error concept. In P. Biemer et al. (Eds.), *Total survey error in practice*. John Wiley & Sons. https://doi.org/10.1002/9781119041702.ch1

Meyer, B. D., Mok, W. K. C., & Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, *29*(4), 199–226. https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.29.4.199

Mneimneh, Z., Lyberg, L., Sharma, S., Vyas, M., Sathe, D. B., Malter, F., & Altwaijri, Y. (2019). Case studies on monitoring interviewer behavior in international and multinational surveys. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop & B. Dorer (Eds.), *Advances in comparative survey methods* (pp. 731–770). John Wiley & Sons. https://doi.org/10.1002/9781118884997.ch35

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

Sastry, N., Axinn, W., de Jong, J., Ghimire, D., & Treleaven, E. (2021). *Assessing the opportunities and challenges in the transition to computer-based survey data collection: The role of key software system features* (Report to the World Bank). University of Michigan Library. https://dx.doi.org/10.7302/3772

Seshadri, S. (2019, November 24). What is consumer expenditure survey, and why was its 2017-2018 data withheld? *The Hindu*. https://www.thehindu.com/business/Economy/what-is-consumer-expenditure-survey-and-why-was-its-2017-2018-data-withheld/article61611662.ece

Thissen, M. R., Fisher, C., Barber, L., Sattaluri, S. (2008). *Computer audio-recorded interviewing (CARI): A tool for monitoring field interviewers and improving field data collection*. Proceedings of the International Methodology Symposium, Data Collection: Challenges, Achievements and New Directions, Statistics Canada, Gatineau, Canada. https://www150.statcan.gc.ca/n1/en/pub/11-522-x/2008000/article/10955-eng.pdf?st=OP1VBs28

Üstun, T. B., Chatterji, S., Mechbal, A., & Murray, C. J. L. (2005). Quality assurance in surveys: Standards, guidelines, and procedures. In United Nations' Department of Economic and Social Affairs, Statistics Division (Ed.), *Household surveys in developing and transition countries* (pp. 199–230). United Nations. https://unstats.un.org/unsd/hhsurveys/pdf/Chapter_10.pdf

Vehovar, V., Slavec, A., & Berzelak, N. (2012). Costs and errors in fixed and mobile phone surveys. In L. Gideon (Ed.), *Handbook of survey methodology for the social sciences* (pp. 277–295). Springer. https://doi.org/10.1007/978-1-4614-3876-2_16

Vila, J., Cervera, J., & Carausu, F. (2013). *Quality assessment of the third European quality of life survey* (Report). European Foundation for the Improvement of Living and Working Conditions. http://doc.ukdataservice.ac.uk/doc/7316/mrdoc/pdf/7316_3rd_eqls_external_quality_assessment_report.pdf

Wuyts, C., & Loosveldt, G. (2019). *Quality matrix for the European social survey, round 8*. European Social Survey. https://www.europeansocialsurvey.org/docs/round8/methods/ESS8_quality_matrix.pdf

Madras Institute of Development Studies (MIDS) was founded in 1971 by Malcolm S. Adiseshiah. It studies development problems of India, with a special focus on Tamil Nadu. Since 1977, MIDS is a national institute funded jointly by the Indian Council of Social Science Research and the Government of Tamil Nadu.

Madras Institute of Development Studies
79, Second Main Road, Gandhi Nagar
Adyar, Chennai 600020, India.
Phone: 044 - 2441 1574 / 2589
pub@mids.ac.in | www.mids.ac.in