35

Case Studies on Monitoring Interviewer Behavior in International and Multinational Surveys

Zeina Mneimneh[1], Lars Lyberg[2], Sharan Sharma[1, 3], Mahesh Vyas[4], Dhananjay Bal Sathe[4], Frederic Malter[5], and Yasmin Altwaijri[6]

[1] *Survey Research Center, University of Michigan, Ann Arbor, MI, USA*
[2] *Inizio, Stockholm, Sweden*
[3] *TAM India, Mumbai, India*
[4] *Centre for Monitoring Indian Economy Pvt Ltd., Mumbai, India*
[5] *Max-Planck-Institute for Social Law and Social Policy, Munich, Germany*
[6] *King Faisal Specialized Hospital and Research Center, Riyadh, Kingdom Saudi Arabia*

35.1 Introduction

35.1.1 Background

The volume of and interest in multinational, multiregional, and multicultural (3MC) survey research have increased considerably during the past decade. This increase has put pressure on survey organizations to produce higher quality data at a lower cost. In spite of this demand, quality assurance (QA) and quality control (QC) programs are still not well developed in multinational and international surveys, especially in countries where survey organizations lack the needed financial, methodological, and technological resources and expertise. The problem is compounded when a strong central team that oversees country activities is missing. Such conditions could lead to variation in data quality across countries compromising the comparability and replicability of data [1]. Yet, as more countries develop their technological infrastructures and catch up on technological and methodological advances in survey research (such as the use of computer-administered personal interviews, mobile technologies, and global positioning systems), adopting a targeted and proactive QA and QC approach tailored to the cultural context becomes more feasible.

Lyberg and Stukel [2] discuss the importance of QA and QC programs and procedures at all the phases of the 3MC survey lifecycle. To achieve accurate

Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC), First Edition. Edited by Timothy P. Johnson, Beth-Ellen Pennell, Ineke A.L. Stoop, and Brita Dorer.

and comparable survey data in multinational initiatives that meet the user's requirements, QA and QC measures and approaches need to be applied. Though QA and QC are often used interchangeably, they do not refer to the same processes or procedures. QA refers to the processes or procedures that aim at accomplishing good quality, such as having a program for interviewer training or collecting interviewer performance indicators. QC on the other hand refers to the activities implemented to make sure that QA actually works.

QA and QC need to be applied at three levels: the survey product, the survey process, and the survey organization [3]. While these three levels are interconnected and one affects the quality of the other, in this chapter we focus on process quality and more specifically on interviewing.

There are a number of reasons for focusing on interviewing QA and QC in international and multicultural settings. First, face-to-face interviewer-administered surveys are still the most common mode in many multinational surveys due to limited penetration of other modes and/or the lack of comprehensive, accurate, and reliable frames in many countries. Second, limited attention has been given to process quality, and specifically to the interviewing process in multinational surveys, possibly leading to greater variation, lower comparability, and sometimes even halting of the fieldwork. Third, the distance between the researcher or the data user and the actual field operation in international and multinational surveys often makes subcontracting the interviewing process to local companies necessary. Many of these data collection organizations are market research oriented and may have interviewing cultures which diverge from standards set by the user and may not be fully aware of the error structures associated with interviewing. Fourth, cultural factors such as privacy concerns, social desirability concerns, and wariness to strangers [4] make it difficult to implement more rigorous QC procedures such as interview recording, and increasing the need for tight but noninvasive interviewer monitoring QC procedures. Fifth, and foremost, interviewers could be a major source of survey error contributing to both bias and variance error components [5–8] including nonresponse and measurement error [9]. In terms of variance (the more studied component of interviewer error), the reported estimates of interviewer intraclass correlations (ICC)[1] in face-to-face surveys are small in absolute terms, often around 0.02–0.05, but sometimes much larger, around 0.1 or more [7, 10, 11]. ICC can vary depending on survey topic and degree of interviewer monitoring. In general, these correlations are about the same magnitude or larger than the ICC caused by geographical clustering [12–14]. A face-to-face survey with an average interviewer workload of 45 and an average ICC of 0.03 for a certain estimate exhibits an increase of 132% in

1 The intraclass interviewer correlation is the correlation between responses obtained within an interviewer's assignment.

variance of this estimate due to interviewer effects.[2] Moreover, interviewer effects could vary across multicultural surveys, thereby reducing comparability. Unfortunately the magnitude of interviewer effect variation has not been thoroughly investigated in multicultural surveys, and there are only a handful of studies that measure interviewer variance on contact and cooperation rates [15, 16], interview length [17], interviewer observations related to respondent's level of effort [18], response easiness [18], interview privacy [19], and substantive topics [11]. Some of these findings show that the magnitude of interviewer variance differs from one country to the other.

In general, interviewer effects occur because of the interactional dynamics between the respondent and the interviewer caused by interviewer characteristics (such as age, gender, race, and religious appearance), because of lack of standardization across interviewers, or because of falsification. The latter two, lack of standardization and falsification, could be triggered by several factors including bad work ethics, inadequate interviewer remuneration, insufficient training, lack of knowledgeable survey managers, or external factors such as bad weather, unsafe neighborhoods, or difficult-to-reach areas (which induce certain interviewers to deviate from the study protocols) [20]. It is easy to imagine how these factors, in addition to the availability of qualified interviewers, could differ from one country to another, causing variation in the magnitude of interviewer effects across surveys. Prevention of such unintentional or intentional deviations from study protocols is mainly attempted through careful interviewer training, appropriate remuneration, and supervision methods such as recording and evaluating interviews, recontacting a subset of respondents to verify the information recorded by the interviewer, or observing the interviewer behavior in the field [21]. The American Association for Public Opinion Research (AAPOR) recommends that a random 5–15% of each interviewer's work is verified or observed [21]. Though increasing this percentage allows for better coverage, it is costly and demands a large team of verifiers if recontacting respondents is to be completed shortly after the interview (say, within two weeks), especially when interviewers' productivity is at its peak. The same applies for increasing the rate of evaluation of recorded interviews and field observations. Moreover, most verifications, observations, and evaluations of recorded interviews cover part of the interview. To augment these procedures (which are typically conducted on a random sample) and target a larger sample of cases worked by specific interviews that require more supervision and evaluation, researchers and survey practitioners have started using data-driven procedures. These data-driven approaches rely on using computer-administered interviews where real-time questionnaire data and paradata (or process data) [22, 23] are analyzed to identify interviewers who exhibit certain outlying behaviors on quality indicators (discussed below) and who require further QC interventions.

2 Interviewer design effect is $1 + (m - 1) \times \text{ICC}$, where m is the average interviewer workload [8, 15].

35.1.2 Using Interviewer-level Data for QA and QC

The use of substantive data and paradata during data collection to guide interventions that aim at reducing error relative to cost is not new to the field of survey research [24–27]. Several terminologies have been used to refer to such designs including adaptive design and responsive design. While the main focus is to intervene at a case level and change design features for certain respondents such as the data collection mode, respondent incentive, extended contact attempts, and response messages (such as a message to slow down while answering), such designs could also be used to refer to targeted *interviewer-level* QC interventions. Under this approach, substantive data and paradata are examined at an interviewer level to guide follow-up and more targeted QC interventions with the objective of creating a more efficient QC system. QC resources are thus channeled to potentially troublesome cases and those interviewers who could be contributing to the majority of the interviewer error.

Implementing such a targeted QC approach begins with identifying a set of critical-to-quality indicators. These indicators need to be proxy measures of potential survey error, whose variation would affect the data quality. They also need to be timely, that is, they could be generated, compiled, analyzed, and intervened on in "real time", shortly after an interviewer has completed the interview. Researchers interested in detecting falsification have used a number of these indicators including household ineligibility rate, number of rare or unlikely response combinations, responses to filter questions, a large number of completed interviews during a short period (say, on the same day), a short interview or question administration time, and the rate of refusal to record the interview [21, 28–34]. While several indicators have been proposed, not many of them have been validated, and further empirical evidence is needed on the association of these indicators with survey errors, especially in multicultural surveys. In fact, the use of real-time interviewer-level indicators to guide QC interventions in 3MC surveys is still very limited in practice. 3MC surveys have started to strive for a more concurrent QC system, but in surveys comprising 25–100+ countries where the technological advances and human resource expertise might be lacking in some countries, disseminating and using such an integrated QC approach across all countries might be difficult to achieve.

Many QC indicators (similar to the ones described above) are based on data residing in multiple databases (interviewing data, sample management system data, keystroke data, etc.). Thus, creating an integrated targeted QC approach that is based on multiple metrics requires devising a system and associated processes that pool these data, process them, automate flagging rules to identify potential troublesome cases or interviewers, and display the outcome. Such systems typically have a multidimensional hierarchical database for better diagnosis of the sources of the problem. The visual display of the output is also an important feature of such a system. The data need to be displayed in a

manner that is easy for QC personnel to visualize, identify troublesome cases, and decide on an action plan.

Once the indicators, the system, and the processes are designed and implemented, interviewers are compared on the indicators to identify any variability patterns through the use of tools such as control charts, ranking procedures, and multilevel modeling or by deciding on a priori cutoffs from earlier waves of the same survey. Interviewers that display special cause variation (observations outside the specified control limits in a control chart) are then identified [3]. Any special cause variation is investigated to understand whether it is an interviewer behavior problem, a geographical clustering problem that is confounded by an interviewer problem, or a respondent behavior problem. To understand the special cause variation, it is important that multiple indicators and different aspects within the specific cultural survey environment are assessed. In studying interviewer falsification behavior, Bredl et al. [35] highlights the importance of taking into account the specific design and specific conditions of the survey before drawing conclusions. A comprehensive investigation is needed before implementing further interventions against an interviewer to avoid incurring additional unnecessary QC costs and also to avoid demoralizing good interviewers for reasons that are not within their control. If it is determined after a comprehensive investigation that the variation could be due to an interviewer behavior, further interventions are needed. These interventions could include retraining the interviewer, sending a supervisor or a QC person to observe the interviewer's behavior in the field, increasing other QC procedures such as verifying more of the interviewer's cases or evaluating more of their recorded interviews, suspending the interviewer temporarily, or terminating his or her service on the project. If variability is large in general but still within control chart limits, then this is a sign of a process that has a large common cause variation that has its roots in a deficient interview process, either at the design stage, the implementation stage, or both. A common cause variation that is deemed unacceptable is decreased by adjusting the interview process itself through improving, for example, the training, supervision, and questionnaire and not by taking action on individual interviewers. Implementing any intervention also calls for testing its effectiveness. Unfortunately, in international and multicultural surveys, systematically documenting when an intervention is implemented and on whom, and making this information readily available for testing the effectiveness of the intervention and modifying the intervention (if needed) are still lacking.

Since the use of such targeted and integrated QC systems is not well developed in international and multicultural surveys, providing guidance and describing case studies that have used similar approaches or methods for measuring interviewer variance across countries would benefit a large number of surveys. This is valuable at a time when computerization has proliferated the

international survey industry, making it more successful and feasible to adopt a closer yet potentially remote control (in addition to local control) of interviewers' work.

This chapter focuses on five international or multinational case studies and the interviewing QC procedures that were implemented. These case studies were selected to represent different types of design and cultural context (e.g. one-time, panel, cross-cultural). Several of them have used technological innovations or modeling techniques illustrating examples of methods that could be used in multicultural surveys. The first case study is the Consumer Pyramids Survey, which is a panel survey conducted in India. This survey employs a large number of interviewers spread over a large and culturally diverse geographic area requiring close monitoring. Among its innovations is the use of a global positioning system (GPS) to monitor interviewers' movements in the field. The second case study is another panel survey in India, the Television Audience Measurement India, that measures TV viewing.[3] The role of the interviewers in this survey is unique and requires attention as it is mainly focused on recruitment, retention, and motivation of respondents to properly register their viewing and report any third-party solicitation that could affect the respondents' viewing behavior contributing to lower data quality. Among its advanced techniques is the use of multilevel modeling to identify interviewers who may be influencing respondents' viewing. The third case study is the Saudi National Mental Health Survey, which is part of the cross-national World Mental Health Survey Initiative. This survey is conducted in a challenging cultural survey environment that requires strict gender matching in a religiously conservative environment where recruiting and training female interviewers is a major challenge, and recording of interviews is culturally unacceptable. This case study uses an integrated QC system that is based on a combination of substantive data and paradata indicators with an interviewer-level and a case-level drill down system. The fourth case study is a multinational survey, the European Social Survey, that provides an example of implementing QC procedures simultaneously across multiple countries and utilizing multilevel modeling to measure interviewer variability across countries. The fifth case study is the Survey of Health, Ageing and Retirement in Europe. This case study highlights the use of a proactive data-driven interviewer-level monitoring approach in a multinational context. None of the case studies are ideal, and they all have their own limitations including the need to empirically test the association between quality indicators and error sources and to test the effectiveness of targeted intervention. Still, they provide a good starting point for researchers and practitioners interested in designing and implementing targeted and integrated QC processes in an international and multinational context.

---

3 In the course of writing this chapter, this panel has since been replaced by another panel operated by another organization.

35.2 Case Studies

35.2.1 Case Study 1: The Consumer Pyramids Survey

35.2.1.1 Background and Survey Design

The Consumer Pyramids Survey is a longitudinal survey of 158 000 households of India that started in 2008. The 158 000 households contain about 550 000 individuals. The objective of this survey is to measure the economic well-being of households in India and its change over time by collecting information on household income, household expenditure and its detailed distribution, savings, involvement in financial investments, borrowing, and ownership of physical assets. The data collected from the panel survey are used to create priced subscription services, and revenues generated from subscriptions are used to fund the execution of the survey and its growth and development. The survey is conducted by the Centre for Monitoring Indian Economy Pvt. Ltd. (CMIE). CMIE designs the survey, implements it, and processes the collected data using its own financial, human, and technical resources.

The survey is conducted through face-to-face interviews. All households in the panel are surveyed three times a year at regular intervals of four months. One complete execution of the survey over the entire panel of households is called a round. A round consists of four monthly slots. Each monthly slot consists of a fixed set of about 39 500 households to be surveyed during its period. A monthly slot is further divided into four weekly slots, each covering seven or eight days. Each household in the panel is planned to be surveyed during a fixed weekly slot of a monthly slot of a round. The schedule of surveying households within a weekly slot is determined by a weekly plan that is created two days before the beginning of the weekly slot. It lists the precise households, the date of interviews, and the interviewer who will conduct the interviews during the week. The exact date when a household should be interviewed and by whom during the week, however, is flexible. This flexibility allows for accommodating local events such as fairs, strikes, heavy rain and flooding, and law and order problems.

35.2.1.2 Management and Operation Structure

The survey operation is managed by a network of eight branch offices and their 14 representative offices. These are located in all the metropolitan cities of India, such as Delhi, Kolkata, Chennai, Hyderabad, Bangalore, and Ahmedabad, and also in other major cities such as Lucknow and Jaipur. Each branch is headed by a branch manager who is responsible for overseeing operations in its regions including hiring and supervising the survey managers. Survey managers in turn oversee their team of field information officers (FIOs). Each FIO manages the survey of approximately 2500 households. FIOs engage and train a team of field team members (FTMs) (interviewers). Typically, there are about 175 interviewers (or FTMs) working in the field per day.

Interviewers are usually fresh graduates with a master's degree in social work. They are hired and trained by FIOs in survey ethics, respondent rapport and motivation, administration of the questionnaire, software operation, logistics, and communication. All interviewers have to pass an online examination and be certified by their survey manager before they start their work in the field.

All interviewers are equipped with smartphones that have data plans and that are GPS enabled. Each day, interviewers download empty data entry screens for the households that have been assigned to them on that day. These empty screens come preloaded with the appropriate addresses and names of household members. Upon locating the household, the interviewer "signs in" at the doorstep of the household and records the GPS-determined location on the smartphone application. The survey begins after this sign-in. Once the survey is completed, the interviewer "signs out" and records again the GPS-determined location. Then the interviewer uploads the survey data to the central database. Once the data are uploaded, they are automatically deleted from the smartphone and become available for quality assessment. The central database also captures and stores the path of the interviewer in the field through the movement of the GPS signals received from the smartphone and the time spent in the field. Such technical infrastructure enables real-time validation of the data uploaded from the field as described later.

Once an interview is validated in real time and accepted, interviewers are paid for the completed interview. Payment is made weekly through a bank transfer. The rate of payment per interview varies across the country and depends on an estimated cost of execution that includes all incidental expenses in the region of the country where the interview is taking place and the region average wage rate per day for comparable labor time.

35.2.1.3 Monitoring Interviewer Behavior

The geographic spread of the survey over three million square kilometers and the total interviewing team of 200 interviewers require close monitoring. Harsh local settings, adverse climatic conditions, and poor transportation infrastructure combined with economic hardship could create an environment that triggers interviewers to deviate from the study protocol.

To monitor interviewers' work, three different types of quality control procedures are used: (i) verification, (ii) data-driven assessment, and (iii) respondent mailing. Each of these procedures is described below:

1) Verification: The local FIO conducts two types of verifications. First, the FIO calls back a random 8% of the households whose data have been uploaded during the day to verify the execution of the survey. Second, the FIO visits and verifies one household per town in urban regions and one household per rural region per interviewer per round. In both types of

verifications (callbacks and visits), the FIO verifies the interviewer who conducted the survey, the duration of the survey, and the mode of administration and collects information on the interviewer's professional behavior. A limitation of the FIO verifications is that they are conducted by the execution team. Therefore, an independent check is also conducted by an audit team. The audit team selects and visits a randomly selected region and verifies that interviewers conducted the survey in the designated households.

2) Data-driven assessment: Data-driven assessment is conducted by the offline real-time validation (ORV) team. This team validates all the data collected by the FIOs and interviewers in near real time. Three types of data are examined for every household, namely, GPS coordinates, keystroke data, and substantive data.
   a) GPS coordinate assessment: The GPS-enabled phones allow the ORV team to check the movement of the interviewers in the field. The team compares the signed in location uploaded by the interviewer to the expected location specified in the weekly plan and accepts entries only if they match.
   b) Keystroke data assessment: The smartphone application that captures the interview data also captures time stamps, audit trails, navigation from one screen to another, and the time taken at each screen. These data are available to the ORV team immediately after the interviewer uploads the survey data to the central database. ORV studies the logic in movements across screens, the time taken at a screen, and the pauses in entries and judges whether the survey was based on a real interview or whether it is possibly fabricated. By the time the interviewer completes interviewing the next household, the previous household's entries are checked, and the results are relayed to the interviewer. The interviewer needs to respond to the ORV's messages before interviewing the next household. This real-time checking provides an opportunity to correct possible errors while the interviewing team is still in the field.

      The process of detecting fictitious entries at this level is manual and judgmental. However, in addition to the real-time evaluation done by the ORV on each interview, the data are also automatically processed the next day, and a week later interviews are pooled together to identify any patterns that could indicate a problem in execution.
   c) Substantive data assessment: Households assigned to interviewers are not constant over rounds. Thus interviewers have no recollection or access to any historical data collected on the households they are interviewing in the current wave. However, the ORV has access to historical data and compares them with the newly uploaded data. In case of discrepancies, the interviewer is requested to contact the respondent,

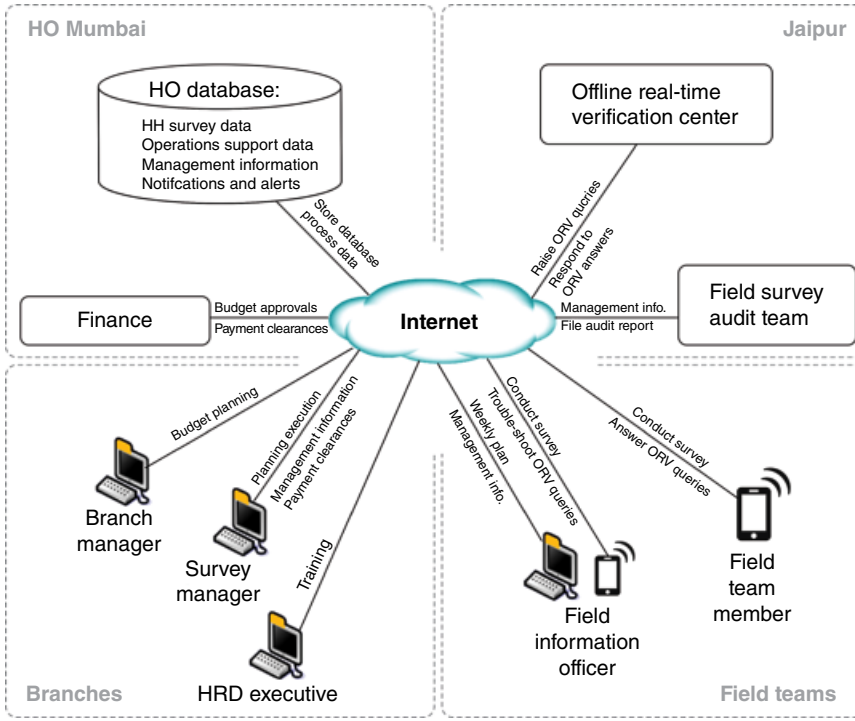


Figure 35.1 Head office, verification center, audit team, field branches, and field teams connectivity. HH, household; HO, head office; HRD, human resource development; ORV, offline real-time verification center.

probe, and clarify the discrepancy in real time while the interviewer is still in the same geographic area. If the respondent confirms the initial response, then it is accepted as a valid observation even if it is not logical or consistent with other observations. The rationale is that economic volatility is expected in India from one wave to the other especially among households that depend largely on farm productivity.

Figure 35.1 summarizes the technical infrastructure that allows the different branches, offices, ORV center, audit team, and field teams that are connected together in real time to collect and verify the data.

3) Respondent mailing: Finally, an annual "thank you" letter is sent from the head office to all respondents thanking them for their cooperation. Letters are delivered to the addressee and acknowledged by the recipient. Letters that cannot be delivered to the addressee are returned to the sender. Returned letters are investigated to check if they reflect a malpractice in the survey execution, a failure of the postal system, or a household that moved.

35.2.1.4 Summary

Real-time monitoring in the Consumer Pyramids Survey is an example of using technology to control for interviewer error and improve the accuracy of the data. The approach is based on the optimal use of real-time information. Every member of the execution team has access to a little more information than the member they supervise. Knowledge that the overall system is continuously monitored, that the management has additional information, and that it uses this information to verify current processes motivates the team to adhere to the survey protocols.

Through the extensive real-time quality controls, the Consumer Pyramids Survey generates data that are ready for estimations in a timely fashion. Everyday verified data for more than 1200 households become available in the central database. The sample is spread across the country. These data can be used to estimate a host of real-time indicators such as daily measures of unemployment and consumer sentiments. Monthly estimates can also be generated from the larger sample of about 39 500 households.

35.2.2 Case Study 2: The Television Audience Measurement Panel in India

35.2.2.1 Background and Survey Design

The Television Audience Measurement (TAM) panel in India started in 1998. The panel comprises about 11 500 households (about 50 000 individuals) across 225 cities in urban India that represent 98% of the estimated TV-owning households in urban India. One of the main objectives of the TAM panel is to estimate TV viewing across channels, which helps advertisers allocate their advertising budgets among these channels. The panel is operationally run by TAM Media Research Pvt. Ltd. (TAM India), a joint venture between AC Nielsen and Kantar Media Research. Funding comes from media organizations such as broadcasters who subscribe to the survey data.

In each selected household[4] viewership data are collected on all individuals four years and older. The data are collected automatically via an electronic device called the Peoplemeter that is attached to the household's TV set. The Peoplemeter captures channel tuning information at the household level, encrypts it, and transmits it wirelessly to the head office for further processing. In order to also capture information on *individual* member viewing, a remote control is provided to each household. The remote has buttons uniquely corresponding to individual members in the household. A viewer presses the

4 Multistage sampling is used to select households. Primary sampling units (PSUs) (cities) are chosen with probability proportional to estimated TV-owning population sizes, while households within PSUs are selected using quota sampling.

button to indicate the commencement of viewing and presses it again to indicate the end of viewing. Data are captured by the Peoplemeter each minute.

The panel is designed to be continuous; data are reported every Wednesday morning for the previous Sunday through Saturday week. Thus, there is only a narrow window available to collect, process, conduct quality control (QC), and report the minute-by-minute viewing data for all 50 000 respondents across more than 600 channels.

35.2.2.2 Interviewers' Role in TAM

Since the data are collected automatically by the Peoplemeter, interviewers play a unique but integral role in TAM. Among the major roles that interviewers play is to motivate respondents to adhere to study protocols. Interviewers visit the assigned households about once a month to remind the members of the importance of button-pushing (i.e. "compliance") and panel security.[5] Interviewers also work with respondents to resolve any participation and compliance issues. For example, the household may perceive that the TV reception has deteriorated after installing the Peoplemeter. Interviewers need to address such concerns to avoid dropouts or noncompliance. In case of dropouts or noncompliance, interviewers are responsible for recruiting a refresher sample.[6]

Another important role interviewers play is updating and collecting household information. A household's profile might change since it was recruited. At the end of each calendar year, the interviewer updates information that was collected from the household when it was recruited. The update interview lasts about 30 minutes and has sections on the household's socioeconomic status, as well as questions related to reception of TV channels, ownership of durables, and lifestyle.

TAM interviewers are paid a fixed salary every month. In addition to the salary, interviewers receive a per-interview payment for the annual household demographic update.

35.2.2.3 Management and Operation Structure

Each state in India has a regional field head (RFH) who supervises a team of interviewers.[7] Given that sample sizes among states vary substantially, the number of interviewers working in each state and reporting to an RFH ranges from 18 to 66. All interviewers (about 530) working on the project are males.

---

5 Household members are reminded to immediately inform the interviewer if they were contacted by any external party with regard to their participation on the panel.

6 The refresher sample also compensates for household removal after serving the maximum time allowed on the panel (three and a half years). The annual turnover rate is 25%.

7 In large self-representing primary sampling units (PSUs), interviewers report to the field head of the PSU who reports to the RFH of the state that contains the PSU.

RFHs are responsible for the panel quality in their state (or self-representing PSU), guided by the national field head. The national field head is assisted by a team of five individuals at the head office who review weekly QC reports and follow up with the state field heads. Twice a year, an annual field conference is held for all RFHs to discuss best practices and conduct refresher trainings. The RFHs in turn conduct refresher trainings to those interviewers who report to them. In addition, a formal interviewer rating system based on quality measures discussed later is updated monthly to provide interviewers with regular feedback on their performances.

35.2.2.4 Monitoring Interviewer Behavior

Interviewers' work in TAM is monitored through field verification and data-driven assessment. Two types of verifications are conducted: internal and external. For internal verification, the RFH visits the panel households assigned to them once or twice a year. Newly recruited households are visited twice a year, once to evaluate the quality of recruitment and once to evaluate the relationship building with the household. External verification is conducted by a top four consulting firm under the direction of the "Measurement Science" (MSci) department, targeting about 1000 households annually (approximately 9% of the sample). The audit sample has a random component as well as a component to purposively include households whose data show unusual viewing behavior. Selected households are visited face-to-face to verify the household demographics and whether the household received the promised incentive on time. During the visit, verifiers also assess the panelists' level of motivation, their relationship with the interviewer, the recruitment process, and whether there have been any deviations from the study protocol by the interviewer or any TAM India personnel. The visit is also important for evaluating whether the respondent is adhering to the study protocol, from recalling their TV buttons and how to press them properly to asking for the identification of any field personnel visiting the household and responding neutrally to any external party inquiring about their panelist status.

While verification is completed on a small subsample, data-driven assessment is done for all households. The continuous upload of data through the Peoplemeter allows for an extensive assessment using both paradata and substantive data.

Since interviewers in TAM play a major role in recruiting, motivating, and coaching respondents on the proper use of the Peoplemeter, the main paradata-driven measures they are monitored on are recruitment rates and nonresponse rates.

Both overall recruitment rate[8] and demographic-specific rates are required to be monitored. Maintaining the overall target recruitment rate is important

8 The overall recruitment rate is the difference between the recruitment target assigned to the interviewer and the number of panel households active under the interviewer.

for achieving the desired sample size. However, an interviewer can meet the overall recruitment target, but the workload could be imbalanced in relation to the target demographic strata. Therefore, recruitment differentials are also reviewed across demographic strata to identify any panel imbalance related to interviewer behavior.

To monitor and minimize interviewers' contributions to potential nonresponse error, interviewer-level unit nonresponse rate is computed and monitored. In addition, a panel turnover rate due to household dropouts is calculated. High turnover rates can suggest dissatisfaction with panel membership possibly related to interviewer behavior. This rate is reviewed by the MSci department every quarter, and the field head office is informed for appropriate interventions.

A household that is successfully recruited into the panel can still contribute to missing data, due to intentional or unintentional respondent behavior. This behavior in turn could be linked to improper coaching on the use of Peoplemeter by the interviewer or insufficient level of motivation. For example, a household might not switch on the TV set during a given analysis period. While this could reflect an actual phenomenon (such as the household being on vacation), it might also be due to a household that intentionally disconnected the meter from the TV set to save them from pressing buttons or due to a perception that the Peoplemeter equipment is increasing their power bill. Missing data could also occur at a household member level where a specific member may have stopped pressing their allotted buttons. This may be due to response fatigue or lack of feedback from the interviewer. These missing data rates are reviewed weekly by the field department, and extreme cases are flagged for further auditing.

In addition to examining the above paradata, substantive data are routinely monitored to identify any unusual patterns. Such patterns could be due to respondents or interviewers being illegally approached by certain entities to influence viewing behavior, thereby biasing the data. An in-house software is used to identify panelists with unusual viewing patterns. Analysts further investigate these members and decide whether to exclude them from the weekly viewing estimates or not. If the unusual pattern continues for more than two weeks, a field visit is made to the household to investigate any possible illegal influence on the panelists' viewing. This process is very labor intensive. To overcome this burden, multilevel models are used to objectively isolate respondent behavior from interviewer behavior and identify interviewers who might not be adhering to the project protocols. The section below describes how multilevel models were used to identify unusual observed patterns.

35.2.2.5 Multilevel Case Study Identifying Potential Falsification

Figure 35.2 displays the viewership trend across 11 consecutive weeks of a particular year for three channels – channel XYZ and two of its competitors

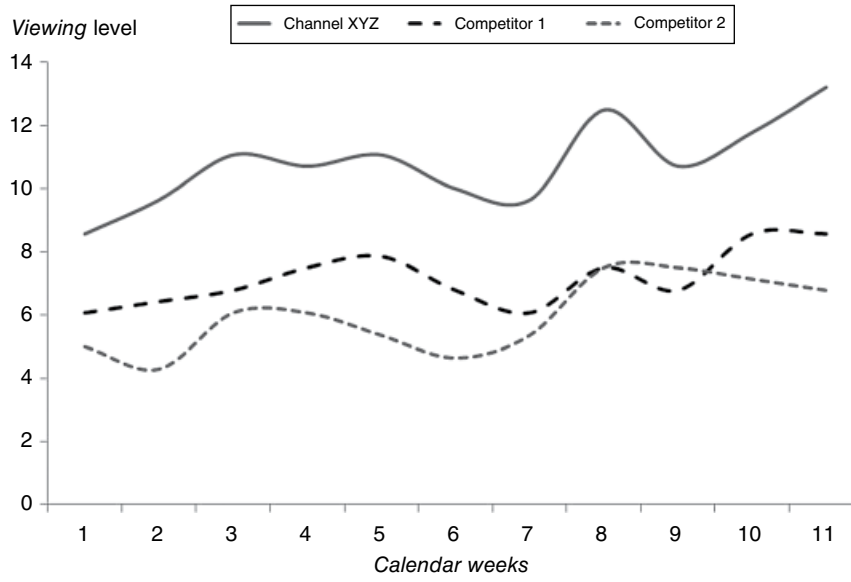


Figure 35.2 Viewership trend for three competitor channels in market A.

in a particular geographic market "A." As the weeks progress, channel XYZ shows a gradual steady increase in its viewing levels, while its competitors have largely maintained the same levels across the same weeks. A further investigation showed that the content of channel XYZ was largely unchanged during this period.

To identify whether this phenomenon could be attributed to specific interviewers, a three-level random effects model was used, with viewing data across weeks nested within households, in turn nested within interviewers. To account for differential household profiles across interviewer workloads, the model also controlled for household characteristics such as socioeconomic status, household size, number of teenagers in the house, and working status of the housewife. The empirical best linear unbiased predictors (EBLUPs) for the interviewer-level random intercepts were plotted as shown in Figure 35.3 (the bold solid lines for each EBLUP represent nominal 95% prediction intervals, while the lighter solid lines represent those based on the Bonferroni correction).

Of the 19 interviewers under study, a visual inspection of the EBLUP plot indicated that interviewers 323 and 319 were outliers.

Drilling down to channel XYZ data of households assigned to these interviewers shows that viewing levels for respondents assigned to the outlier interviewers are unusually large as compared with those assigned to other interviewers. It is important to note that channel XYZ content did not change across weeks, and even if there was a renewed interest in some of the same content,

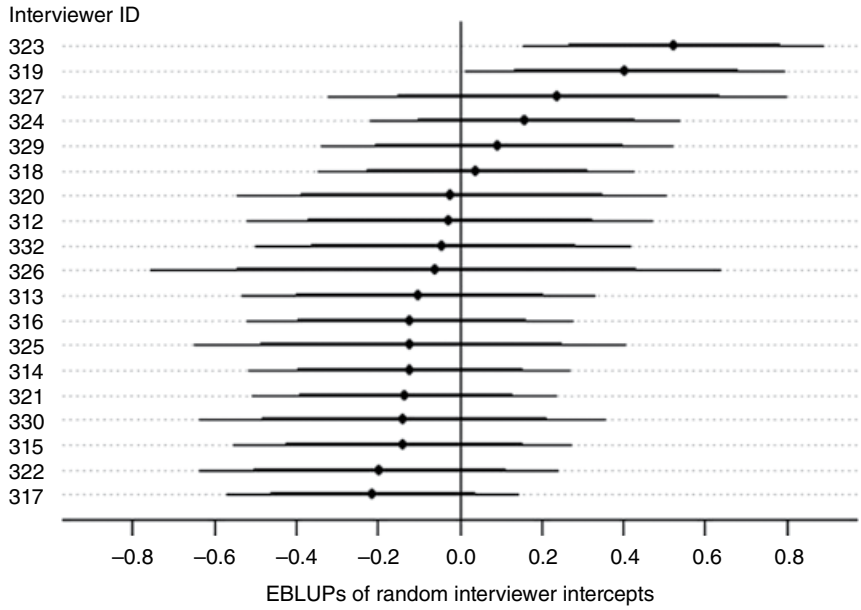


Figure 35.3 Interviewer-level random intercepts (EBLUPs) for channel XYZ.

one would expect this to hold for households across interviewers. However, only two of the 19 interviewers account for the heavy viewing households. Though this analysis identifies the outlying interviewers, it does not confirm that this behavior is due to interviewer falsification, i.e. in this case interviewers asking their respondents to watch channel XYZ. It is possible that channel XYZ conducted heavy promotions in the areas controlled by these two interviewers. In such cases, further evaluation and a comprehensive assessment helps us better understand the reasons behind what seems to be suspicious data. However, pending further investigation, the heavy viewing households assigned to the two interviewers were withheld from further reporting.

35.2.2.6 Summary

TAM India implements both traditional weekly quality control methods and analyses and model-based QC methods. The model-based methods are run once every month and are interpreted in conjunction with the traditional weekly analyses to identify interviewers or household behaviors that require an intervention. The combination of these methods has shown great promise in saving time and detecting outlying behaviors at the household and interviewer level, guiding further interventions. Large-scale multinational studies could benefit from such model-based methods where country variation could be estimated once enough interviews are conducted in each country.

35.2.3 Case Study 3: The Saudi National Mental Health Survey

35.2.3.1 Background and Survey Design

The Saudi National Mental Health Survey (SNMHS) is part of the World Mental Health (WMH) cross-national initiative. The WMH initiative is comprised of more than 30 community surveys conducted across the world [36]. SNMHS is conducted and funded by several local Saudi organizations and methodologically supported by the two WMH coordinating centers at Harvard University (WMH Data Analysis Center) and the Institute for Social Research at the University of Michigan (WMH Data Collection Coordination Center). The survey aims at estimating the prevalence of mental health disorders in the Kingdom of Saudi Arabia and investigating their risk factors, burden, and treatment. The survey has been in the field since 2013 with long interruptions due to weather conditions (the field is halted during the summer season), funding changes, and field operation changes.

SNMHS is based on a national multistage area probability sample of 5000 households from the 13 administrative areas of the Kingdom. Within each selected eligible household, a male and a female Saudi between the ages of 15 and 65 are selected randomly. Interviews with selected respondents are gender-matched and are conducted face-to-face using the Saudi version of the Composite International Diagnostic Interview (CIDI 3.0) [36, 37]. Computer-assisted personal interview (CAPI) is used with audio-assisted components (ACASI) for sections asking about sensitive information such as suicidal behavior, religiosity, and alcohol and drug use. Respondents are also asked to give saliva samples at the end of the interview.

Collected data are sent daily to a central server in Ann Arbor, Michigan. Interviewers send and receive data using a University of Michigan in-house sample management system. Interviewers are trained over two weeks. The training covers general interviewing techniques, CIDI-specific training, hardware and software use, ACASI administration, saliva collection, sample management system features, and data collection protocols. All interviewers have to pass a face-to-face certification before starting their fieldwork. Interviewers are remunerated based on a combined per-case and per-hour structure. Completed main interviews are paid per hour, whereas final main noninterviews and household screener interviewers are paid per case. Interviewers are also compensated for transportation cost, phone and Internet charges, and accommodation (for the travel teams). In addition, a monthly bonus is given to interviewers who complete a high number of main interviews that pass the quality checks described later.

35.2.3.2 Management and Operation Structure

The first wave of data collection was contracted to a private local company. After that, the fieldwork was handed back to the local research team. Since

then the local research team has hired and trained its own project staff, interviewers, and supervisors. The project staff consists of one project manager, two field managers, one quality control (QC) coordinator, four QC staff, four help desk staff, and one data manager. Most of the interviewers are hired locally from each region and are supervised by a group of team leaders at an overall ratio of 4 : 1. The local interviewing team is supported by a small traveling team that works across regions. The fieldwork is designed to be staggered across regions to maintain close supervision. At no point in time are there more than 30 interviewers active in the field.

In addition to the close supervision by the team leaders, interviewers are monitored through a combined approach using routine QC procedures (verification and field observation) and real-time data-driven assessment. This extensive approach is needed in the Saudi survey culture as there are many factors that could drive interviewers to take shortcuts or falsify data. Such factors include the harsh climate, paucity of complex academic face-to-face surveys, wariness of the Saudi population to strangers visiting their households, and the absence of interview recording.

35.2.3.3 Monitoring Interviewer Behavior

Two types of routine QC procedures, verification and field observation, are used to evaluate interviewers' work in SNMHS. Verification is conducted on a random 10% of completed interviews and 5% of noninterviews. The majority of verifications are conducted by telephone within two weeks from the time the interview is completed. Occasionally, when the selected household does not have a phone number or when it is not possible to establish contact with the household (i.e. no one answers the phone), verification is conducted face-to-face. Verifiers follow a script that confirms whether an interviewer visited the house and the final outcome of the respondent–interviewer interaction. For completed interviews, verifiers also administer a set of questions from those interviews and confirm the ACASI administration and the saliva sample request.

Field observations are conducted by team leaders. During field observations, team leaders accompany their interviewers and evaluate them on their adherence to the survey protocol and their interactions with respondents, an important aspect of interviewing that cannot be captured during verification. An initial evaluation is conducted for all interviewers within their first two weeks of production.

In addition to the random and initial cases that are selected for verification and field observations, interviewers can be flagged for additional verification or field observation. Targeted interviewers are identified based on the results of the real-time data-driven assessment of all completed interviews and noninterviews.

For data-driven assessment, a set of interviewer-level QC indicators is generated in real time from both paradata and substantive data. These indicators are

classified into two groups: *single* occurrence indicators and *cumulated* indicators. Single occurrence indicators are the timeliest and require an immediate intervention. Interviewers are flagged on any of these indicators whenever they have completed a case that exceeds a specific cutoff on an indicator. An example of such indicators is "question not read." The cutoff for this indicator is one second or less, indicating a question that potentially was not read by the interviewer. The second group of indicators, cumulated indicators, aims at detecting a potential pattern and requires accumulation of completed cases over a certain period of time depending on the interviewer productivity. Each interviewer is compared with the rest of the interviewers on these indicators, and extreme cases are detected. Examples of these indicators include the average interview length and eligibility rates. Interviewers who have the lowest three averages or rates are flagged.

Within each of these two groups, indicators are also classified by the potential error type they can be associated with. At this stage, the error classification is not based on empirical evidence but on a hypothesized relationship between the indicator and the potential error. Additional analyses are needed to establish potential associations between these indicators and the relevant error type.

Table 35.1 summarizes the quality indicators used in the SNMHS and their classification into single versus cumulated indicators and the error source.

Data underlying all quality indicators are stored in multiple databases (e.g. audit trail, sample management, and questionnaire data) and are compiled and displayed daily in a tool called the Audit Trail Online Analytical Processing (ADT OLAP) cube. The ADT OLAP cube makes it possible to have all these data readily available to managers who can manipulate them as needed via Excel pivot functionality. The OLAP cube displays the indicators by each interviewer. A summary sheet is available that shows the current status of all single occurrence indicators and cumulated indicators by interviewer. Table 35.2 is an example showing the flagged status (1 flagged, 0 not flagged) of each single occurrence indicator by interviewer.

In addition, a detailed table for each indicator is provided. Detailed tables allow drilling down to the specific date, specific interview, and specific question that triggered the flag. Figure 35.4 is an example of "question not read" report showing interviewers' flagged status and the drilling capability of the OLAP cube by date, sample ID, and question field name.

Quality indicator tables are typically reviewed by the local Saudi QC coordinator, and recommendations for interventions are discussed with the project manager. The frequency of reviewing the indicators depends on their type. Single occurrence indicators (such as "question not read" during the interview) are reviewed daily, while cumulated indicators are reviewed weekly or biweekly depending on fieldwork productivity. The longer period for cumulated measures is needed to allow for the detection of any change in interviewers' behavior over time. Below is a summary of the types of comprehensive follow-up actions that can be triggered by flagged quality indicators.

Table 35.1 SNMHS quality control indicators by sources of errors.

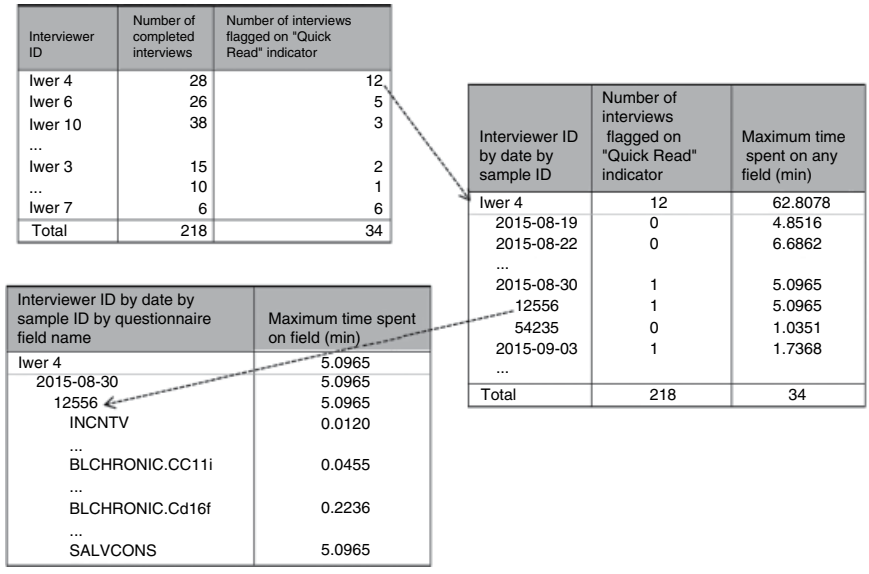
Source of error	Single occurrence indicator (cutoff)	Cumulated indicator (highest or lowest three rates: H vs. L)
Measurement	• A pause during an interview (10 min or longer)	• Rate of cases unable to verify (H)
	• An interview with a question read too fast or not read at all (1 s)	• Rate of verifications with a discrepant outcome (H)
	• A very short interview (less than 30 min)	• Rate of interviews with short path (H)
	• Number of interviews completed on the same day (three interviews)	• Rate of interviews with no mental health disorders (L)
	• Failed verification (one instance)	• Short average interview length (L)
		• Rate of interviews switched from ACASI to CAPI administration (H)
Coverage	• Short travel time between two interviews on the same day (10 min)	• Rate of cases unable to verify (H)
	• Deletion of listed household members (three interviews with at least one deletion)	• Rate of households with no eligible female (H)
	• Failed verification (one instance)	• Rate of households with no eligible male (H)
Nonresponse		• Rate of saliva refusals (H)
		• Average contact attempts per completed interview (L)
		• Response rate (L)

H, highest three rates; L, lowest three rates.

Flagged single occurrence indicators typically require checking contact history, contact notes, or question field notes recorded by the interviewer, and gathering information on the flagged case from the field help desk, team leader, interviewer, or sometimes the respondents. For example, if a completed interview has a "question not read" flag, the QC coordinator will check any notes made by the interviewer that could explain this behavior. If no information is found, the help desk, team leader, and interviewer are contacted to collect further information and to get an understanding of the specifics of the situation. Respondents may also be contacted by telephone or face-to-face to

Table 35.2 Example of a summary sheet showing the flag status (1 flagged, 0 not flagged) for each interviewer on single occurrence indicators.

Interviewer ID	Count of completed main interviews	Question not read	Failed verification	Short interview	Long pause	Deleted household member	Number of completed interviews/ day	Short time between interviews	Sum of flagged indicators
1	6	0	0	1	1	0	0	0	2
2	2	1	1	0	1	0	0	0	3
3	13	1	0	1	1	0	0	0	3
4	15	1	0	1	1	0	0	1	4
5	69	1	0	0	1	0	0	1	3
6	29	1	0	1	1	0	1	1	5
7	26	1	0	0	1	0	1	0	3
8	38	1	0	0	1	0	0	1	3
9	16	1	0	1	1	0	0	1	4
10	27	1	0	1	1	0	0	0	3
…	…	…	…	…	…	…	…	…	…
Etc.									



Interviewer ID	Number of completed interviews	Number of interviews flagged on "Quick Read" indicator
Iwer 4	28	12
Iwer 6	26	5
Iwer 10	38	3
...		
Iwer 3	15	2
...	10	1
Iwer 7	6	6
Total	218	34

Interviewer ID by date by sample ID	Number of interviews flagged on "Quick Read" indicator	Maximum time spent on any field (min)
Iwer 4	12	62.8078
2015-08-19	0	4.8516
2015-08-22	0	6.6862
...		
2015-08-30	1	5.0965
12556	1	5.0965
54235	0	1.0351
2015-09-03	1	1.7368
...		
Total	218	34

Interviewer ID by date by sample ID by questionnaire field name	Maximum time spent on field (min)
Iwer 4	5.0965
2015-08-30	5.0965
12556	5.0965
INCNTV	0.0120
...	
BLCHRONIC.CC11i	0.0455
...	
BLCHRONIC.Cd16f	0.2236
...	
SALVCONS	5.0965

Figure 35.4 Example of "question not read" report showing interviewers flagged on any single occurrence of "not read" indicator and the drilling capability by date, sample ID, and question field.

confirm certain information. A decision is then made on whether this flag is triggered by a respondent behavior, an interviewer behavior resulting from lack of protocol knowledge, or an interviewer behavior that is a deliberate deviation from the protocol.

Flagged cumulated measures typically require checking or observing multiple cases to interpret potential patterns. For example, if an interviewer ranks high on the rate of obtaining ineligible female respondents, households are recontacted to investigate whether the lack of eligible females is a real phenomenon or whether the interviewer manipulated the household roster. Interviewers may also be accompanied by their team leader to observe how they are handling the household roster and to discern whether the eligibility pattern is due to insufficient training or not. Certain flagged indicators might necessitate multiple field observations over an extended period of time (e.g. for five completed interviews) to confirm whether the pattern observed is due to interviewers' intentional deviations from the protocol. Such conclusions can be made when the data collected during field observations differ from data collected without observations, possibly indicating a change in interviewer behavior when they are closely supervised.

Some interventions require observing patterns across multiple indicators (rather than a single indicator) or observing a change in indicator over time.

For example, an interviewer flagged because of recent low endorsement rates of "gate"[9] questions may be due to the assigned primary selection unit or an interviewer's better understanding of the instrument (and thus taking shortcuts). To isolate these effects, the indicator is monitored over time.

Depending on the result of the follow-up actions and the observations collected during them, any of the following actions may be needed: (i) No definitive deviation from the protocol has been identified, so monitoring the interviewer behavior is continued; (ii) retrain the interviewer on a specific component of the study such as general interviewing techniques, administering the household roster, or getting saliva consent; (iii) suspend the interviewer for a period of time until a further investigation on the interviewer's work is conducted; or (iv) remove the interviewer from the study permanently.

At the time of writing, data on the occurrence and the type of intervention were not merged to empirically test the effectiveness of the intervention. However, based on the weekly–biweekly discussions with the local project manager, retraining some of the interviewers on a specific component of the study, such as saliva administration, helped improve the corresponding quality indicator. Moreover, though there has not been any incidence of confirmed falsification, a number of interviewers who were regularly flagged, and whose work was questioned, were suspended temporarily. These interviewers left the project voluntary after they experienced these tight QC processes. Thus these QC processes seem to have deterred interviewers from taking shortcuts and therefore collecting lower quality data.

35.2.3.4 Summary

Developing QC processes and procedures that combine the more traditional routine methods, such as verification and field observation, on a random subsample with real-time monitoring of data and paradata has proven to be useful for the SNMHS. The data-driven approach helped target certain cases or interviewers for additional QC, creating a more efficient QC system. The need for such an integrated approach is especially important when interviews are not audio-recorded. However, implementing these processes required up-front setup and training for local staff. Moreover, several areas of improvements are needed to make the system more effective and efficient. These improvements include reducing the number of indicators and potentially clustering them into factors, using more color coding and symbols for usability purposes, using statistical QC charts instead of ranking (since ranking is sensitive to workload), and most importantly automatically linking the different QC procedures together and establishing a dynamic integrated adjustment to the processes, a

9 A gate question is a question that if endorsed will lead to asking a series of other questions and if not endorsed will lead to a shorter interview path.

procedure that is now being integrated across other projects at the Survey Research Operations at the University of Michigan. Finally, further work is needed on empirically testing the association between the implemented set of quality indicators and survey error and the effectiveness of the interventions tied to these quality indicators.

35.2.4 Case Study 4: The European Social Survey

35.2.4.1 Background and Study Design

The European Social Survey (ESS) is an international comparative survey conducted every two years since 2002 (www.europeansocialsurvey.org). It was initiated to fill a scientific interest in social science by academics and a political and governance interest by the European Commission. It is primarily an attitude survey, albeit not exclusively. The number of participating countries varies across rounds, and so far 36 countries have taken part in at least one round, including 27 European Union (EU) states (not Malta), Norway, Switzerland, Israel, Turkey, Russia, Ukraine, Iceland, Kosovo, and Albania. Nationally representative samples of individuals are selected in each participating country using probability sampling, and questionnaires containing repeat and rotating modules are administered in approximately one hour face-to-face interviews.

There are three main parts of the ESS lifecycle. First, questionnaire development starts with a Europe-wide competition for question module design followed by module development, pretesting, and final translations using the TRAPD team translation model (draft translation, review and refine sessions, adjudication for pretest, pretest, and ongoing documentation) [38], linguistic verification, and Survey Quality Predictor (SQP) coding [39]. The second part is the implementation process, which includes sampling, data collection, and data processing. Since the data collection mode is uniformly face-to-face, hiring and training of interviewers are crucial. Interviewers in all participating countries are expected to have received interviewer training and to have experience with face-to-face surveys in a probability sample survey setting. Interviewers with experience only in nonprobability approaches such as quota sampling are required to have more extensive training covering respondent selection protocols before being considered as ESS interviewers. Interviewers are also trained on implementing procedures to enhance response rates, including contact strategies, doorstep interactions, refusal avoidance and conversion techniques, and reissuing refusals and noncontacts.

Special efforts are made to encourage interviewers to achieve reasonable response rates, especially among sample members that are hard to reach. Techniques for maximizing response vary by country. In the United Kingdom, for instance, interviewers are told not to call the survey ESS but rather the Living in England/Wales/Scotland/Northern Ireland Today Survey. The reason

is that many British people are anti-EU, which was confirmed by the outcome of the 2016 referendum on UK EU membership. However, when delivering the survey request, the interviewer might sense a sampled person's interest in comparative aspects, which might fit within the commitment and consistency compliance principle. Interviewers are encouraged, when deemed suitable, to use this and other compliance principles described in Groves and Couper [40] to motivate people to participate in the survey.

The third part of the ESS lifecycle is quality assessment and documentation. Quality assessment is implemented across the different survey phases. The sample designs are signed off by a central expert team. Translations are subject to linguistic verification by an external service provider [41] and check of formal characteristics by using the SQP coding. Adherence to other specifications is assessed albeit late in the processes. National coordinators who lead the work in each country are required to provide quality reports including information on interviewer selection, interviewer experience, description of training activities, payment model, and number of interviews back-checked.

35.2.4.2 Management and Operation Structure

The main central funding for rounds 1–6 of the ESS came from the European Commission. However, over the years funding has come from multiple sources including the European Science Foundation. Participating countries have obtained funding to cover large portions of their own data collection and survey management from their national research foundations and government entities. In 2013 the ESS became a European Research Infrastructure Consortium (ERIC) where participating countries contribute in varying degrees to the financing of the central activities with some supplementary support from the European Commission.

The organizational structure of ESS consists of a number of expert panels and multinational committees. The ESS-ERIC is governed by a General Assembly that appoints the Core Scientific Team (CST). CST is comprised of a number of groups that are responsible for certain activities, such as developing questionnaire modules, pretesting, translation, sampling, and providing advice on methods and contents.

According to the statutes of the ESS-ERIC, each country (including non-member participating countries) shall appoint a national coordinator who is in charge of the country survey. The national coordinator is in charge of the local planning and helps select a suitable local survey organization. The coordinator is responsible for discussing the ESS specifications with the local survey organization and for leading the training and briefing of fieldwork staff. The survey organizations are responsible for hiring and compensating their interviewers. Survey organizations vary in their interviewer payment structure from a fixed salary, hourly payment, to per-completed interview payment. Bonuses and specific fees for working on a set of sample units (like initial refusals) are

sometimes provided as well. Though the local survey organization is in charge of the data collection, the national coordinator is responsible for overseeing fieldwork, and he or she works closely with the survey organization to provide weekly or biweekly fieldwork progress reports to the CST. Some countries opt to select a coordinating team rather than a single coordinator.

35.2.4.3 Monitoring Interviewer Behavior

There is a process for interviewer quality monitoring in ESS, albeit not an ideal one. The survey organization monitors interviewers. The national coordinator monitors the survey organization using fieldwork reports. The CST's country contact oversees the entire process and monitors information on contact attempts, noncontacts, and response rates. Identification of deviations from specifications and analyses of interviewer effects are conducted after the fieldwork has been completed and often serve as input to methodological reports by members of the CST.

This case study summarizes (i) the quality control procedures implemented by the local survey organizations conducting the country-specific surveys and most importantly (ii) the procedures used to investigate interviewer variation in data quality across countries.

Interviewers' work in ESS is locally monitored through the use of routine control procedures including back-checks and field reports.

A subsample of each interviewer's work is selected for back-check. Back-check is conducted on 10% of each interviewer's completed interviews and 5% for noncompleted interviews including noncontacts, refusals, and ineligibles. Respondents are contacted by telephone or face-to-face and are asked whether the interview actually took place, the type of questions that were asked, whether show cards were used, and the approximate length of the interview.

In addition to back-checks, the survey organization in each country provides the national coordinator with periodic response rate[10] reports by regional levels, respondent subgroups, and interviewer. The reports also include information on the average length of interview for each interviewer. The survey organization investigates any interviewer who displays an outlier pattern on this metric. In order to provide those reports and investigate any outliers, all interviewers are required to complete a contact form for each household assigned to them. On those forms the interviewer records his or her identification number, the household and respondent selection procedure, the date, time and outcome of all contact attempts (visits or calls in some countries), demographic information of the initial refusers, the interviewer's judgment of future cooperation for initial refusers, and information on dwellings and neighborhoods.

---

10 The response rate target is 70% percent although the achieved response rate can be lower in some countries. The proportion of noncontact rates should not exceed 3% of all sample units.

Data collected on these forms are also used to check whether interviewers are making at least four contacts to each sample unit before they classify the unit as incomplete and whether the contacts are made on different days and times of the day and spread over two weeks.

Contact forms also play a central role in investigating the quality of the data collected across countries. All completed contact forms are available at www.europeansocialsurvey.org (fieldwork documentation) and can be used to identify any unusual interviewer variation across countries. Unfortunately, there is a time lag between the completion of the forms and the time they are made available, making it difficult to centrally intervene in a timely fashion. However, since the data are collected over different periods in each participating country, analyzing any paradata across countries in real time or close to real time is not possible.

One example of how collected data are analyzed to investigate quality variation across country and interviewers is presented in Table 35.3 [11]. Table 35.3 shows the average intraclass correlation (ICC) over 48 items across 36 countries in ESS. A seemingly small ICC can still increase the variance substantially if interviewer workload is large. This is the reason for the ESS requirement regarding a maximum interviewer workload size of 48. This requirement however is frequently not adhered to [11].

As can be seen from Table 35.3, the intra-interviewer correlation, ICC, varies substantially between countries, and this will affect cross-country comparisons. ICC is normally between 0 and 0.05 in an environment that puts emphasis on controlling interviewer effects [7], but Beullens and Loosveldt [11] show that some countries including Bulgaria, the Czech Republic, Greece, Kosovo, Lithuania, Romania, the Russian Federation, Slovakia, and Ukraine have average ICC values between 0.15 and 0.20 and sometimes even higher. Such high values are not expected to occur if interviewing is done in a standardized fashion and according to specifications. Thus despite detailed instructions issued to all participating countries, such large variations are still observed. As a result of high ICCs, comparability is compromised, and variances are underestimated, sometimes considerably. Thus, interviewer error is probably the most serious error source in 3MC face-to-face surveys, but its characteristics are not well known by most service providers, researchers, and other users of ESS data. Beullens and Loosveldt [11] point out that almost all the authors of the 221 ESS substantive research articles published in 2013 ignore information about interviewer variance and its analytical consequences.

Loosveldt and Beullens [17] have also analyzed variations in interview length in the fifth round of ESS. Much variability has been observed between interviewers and countries. Average interview length ranges from 102 min (Czech Republic) to 51 min (Slovenia). Such differences are unexpected and are an indication of interview practices that deviate from specifications.

Table 35.3 Summary of intra-interviewer correlations over 48 survey items for 36 countries in six ESS rounds.

	Average intra-interviewer correlation					
	1	**2**	**3**	**4**	**5**	**6**
Albania						0.05
Austria		0.09	0.09			
Belgium	0.04	0.06	0.05	0.05	0.05	0.05
Bulgaria			0.22	0.23	0.23	0.24
Croatia				0.11	0.14	
Cyprus			0.13	0.14	0.18	0.15
Czech Republic	0.16	0.18		0.21	0.04	0.26
Denmark	0.03	0.03	0.03	0.03	0.03	0.02
Estonia		0.11	0.12	0.01	0.07	0.10
Finland	0.02	0.02	0.02	0.02	0.02	0.02
France		0.04	0.05	0.04	0.04	0.04
Germany	0.08	0.10	0.13	0.08	0.07	0.05
Greece	0.16	0.19		0.23	0.22	
Hungary	0.06	0.08	0.10	0.12	0.10	0.16
Iceland		0.01				0.01
Ireland	0.10	0.08	0.07	0.05	0.15	0.16
Israel	0.11			0.18	0.15	0.12
Italy						0.07
Kosovo						0.27
Latvia				0.15		
Lithuania					0.16	0.28
Luxembourg		0.09				
The Netherlands	0.02	0.02	0.03	0.03	0.02	0.02
Norway	0.02	0.02	0.03	0.01	0.02	0.02
Poland	0.08	0.09	0.09	0.10	0.11	0.10
Portugal	0.16	0.15	0.19	0.16	0.19	0.13
Romania				0.23		
Russian Federation			0.18	0.22	0.20	0.22
Slovakia		0.08	0.12	0.19	0.17	0.22
Slovenia	0.05	0.03	0.06	0.07	0.09	0.09
Spain	0.15	0.11	0.09	0.13	0.07	0.05
Sweden		0.01	0.02	0.02	0.02	0.07
Switzerland	0.05	0.06	0.06	0.06	0.06	0.05
Turkey		0.15		0.21		
UK	0.03	0.04	0.04	0.04	0.05	0.06
Ukraine		0.21	0.23	0.22	0.24	0.26

Adapted from Buellens and Loosveldt [11].

One could speculate that local fieldwork practices might take precedence over ESS specifications, a finding that suggests the need for closer monitoring to understand such variations.

Finally, there are several limitations that are important to note regarding the quality control procedures used in ESS locally and cross-nationally. First, most of the fieldwork quality control measures used in ESS rely on data collected by interviewers themselves, which is far from optimal, as they could be inaccurate. Second, indicators related to response rates are difficult to interpret. High response rates might reflect good interviewer efforts, skilled interviewers, efficient interviewer-to-supervisor ratios, or many contact attempts. However, they can also reflect aggressive refusal conversion resulting in increased measurement error, an excessive number of contact attempts, or even fabrication of data. More objective quality measures and indicators are needed, and these must be defined with fixed essential survey conditions in mind [42] to be comparable across countries.

35.2.4.4 Summary

The ESS has built a strong infrastructure that will facilitate future improvements, especially after the forthcoming adoption of CAPI in many ESS countries. Country activities as well as central activities would be monitored through indicators currently collected by the interviewers. An overall quality assessment profile could then be created that combines indicators about central activities with country profile indicators for a specific ESS round.

An important aspect of any indicator is its timeliness and the ability to intervene while the data collection is still in progress. Currently all actions based on indicators come late, if at all. With the adoption of CAPI in all countries, more real-time indicators including those related to response patterns, uneven workloads, and interview length will be available, and interventions become possible. Ideally the central team at ESS would then have a global dashboard so that all countries could be followed simultaneously. Such a system would be a logical future improvement.

35.2.5 Case Study 5: The Survey of Health, Ageing and Retirement in Europe (SHARE)

35.2.5.1 Background and Survey Design

The Survey of Health, Ageing and Retirement in Europe (SHARE) is a multidisciplinary cross-national panel study that assesses the health, socioeconomic status, and social and family networks among individuals 50 years or older in 20 European countries and Israel. As of wave 6 (completed in November 2015), more than 123 000 individuals were interviewed (approximately 293 000 interviews) using a computer-assisted personal interview (CAPI) mode. SHARE was initiated in response to a call by the European Commission to assess the

possibility of establishing a European Longitudinal Ageing Survey. SHARE is closely harmonized with its global partner studies such as the Health and Retirement Study (HRS) in the United States. Since 2004, panel data are collected every other year from the same individuals (and from additional refreshment samples accounting for panel attrition).

35.2.5.2 Management and Operation Structure

In March 2011 SHARE became the first European Research Infrastructure Consortium (SHARE-ERIC). SHARE is centrally coordinated by the Munich Center for the Economics of Aging (MEA), within the Max Planck Institute for Social Law and Social Policy. Input (or *ex ante*) harmonization is a key feature of SHARE. All countries are required to use the same questionnaire (translated from English into national languages) and the same software tools. In every country, the implementation of the design is coordinated by a scientific partner organization, the country team, which is usually affiliated with a national university. The actual data collection is subcontracted to mostly for-profit survey agencies. These survey agencies recruit, hire, and manage the interviewers. All survey agencies involved in wave 5 of SHARE pay their interviewers per completed interview. Many of them compensate their interviewers for travel and provide extra incentives for achieving high response rates or converting initial refusals. Some survey agencies also provide extra incentives for reaching a specific number of completed interviews in a given time frame. All survey agencies are required to follow the same timeline and produce the same set of deliverables (interview data and paradata). To enforce this strong ex ante harmonization, SHARE has established an integrated legal and quality control (QC) framework through a set of documents, deliverables, and procedures. A model contract specifies all legal agreements between the survey agency and SHARE (including costs and payments, property rights, data protection, liability, and audits). The contract also specifies the survey design and implementation standards under which SHARE operates. These specifications address eligibility rules and sampling, household contact protocols, interviewer trainings, interviewer payment, minimum retention rate for panel samples, minimum response rate for the refreshment samples, workflow between the survey agency, country teams, and SHARE Central in Munich, and finally all QC measures performed both by the survey agencies that manage the interviewers and by SHARE Central that manages the survey agencies.

35.2.5.3 Monitoring Interviewer Behavior

Interviewers on SHARE are monitored through both routine QC procedures and real-time data-driven assessment. For routine procedures, the survey agency verifies at least 20% of completed interviews. Verification interviews are conducted over the phone and inquire about whether an interviewer visited

the house, the interview duration, the interviewer's professional conduct, and the use of a laptop and dynamometer and also include a set of questions from the actual interview.

For the data-driven approach, starting with wave 5, fieldwork monitoring was reconceptualized with explicit reference to the total survey error (TSE) framework [43]. This strategy helped prioritize fieldwork monitoring with the aim of minimizing various survey error components [44]. The most important innovation was a proactive approach to provide survey agencies with interviewer-level indicators and to request certain interventions for interviewers who were underperforming on selected quality indicators. Quality indicators include the rate of attempted households, the rate of "reached" households (i.e. households where a contact has been established), cooperation rate, refusal rate, median interview length, and reading time of long introductory items in the questionnaire.[11] Each of these indicators has a predefined cutoff (detailed in Table 35.4) for flagging interviewers.

This new approach also requires publishing the outcomes of the QC procedures in a report called "compliance profiles" (e.g. for wave 5 see Ref. [45]).

To implement this data-driven proactive approach, all survey agencies need to install a server that holds all addresses of the sample and that assigns households to interviewer laptops (the "sample distributor" [SD]). The agencies also need to install an integrated software tool containing the actual interview software (programmed in Blaise) and the electronic contact protocol (the "sample management system" ["SMS"] implemented with JavaScript) on all interviewer laptops (leftmost box in Figure 35.5). All data generated by the interviewers including questionnaire data, sample management data, and paradata are synchronized with the agency SD on a daily basis. The data are then sent to CentERdata (SHARE's IT partner located in the Netherlands) biweekly. CentERdata processes the data and sends them to SHARE Central within five working days after the survey agencies export their data. All dates of data exports from agency servers to CentERdata servers are fixed before the fieldwork starts.

The fieldwork monitoring team at SHARE Central then combines the data from all countries and generates biweekly reports with country-level statistics over time. The reports are sent to all country teams and contracted survey agencies. In addition, extensive Excel tables related to quality indicators are generated at the interviewer level. These interviewer-level measures are derived from paradata using automated routines[12] and are then manually edited for better readability. Since each interviewer is assigned a laptop, the

11 For details of the computation of these indicators, please see the methodology volume of wave 5: http://www.share-project.org/fileadmin/pdf_documentation/Method_vol5_31March2015.pdf.

12 For technical details on data extraction, see the methodology volume of wave 5: http://www.share-project.org/fileadmin/pdf_documentation/Method_vol5_31March2015.pdf.

Table 35.4 Laptop-level statistics on quality indicators and flagged underperforming interviewers (in bold) for SHARE wave 6.

Laptop ID (interviewer)	Total number of days in the field	Total number of assigned panel HH[a]	Rate of attempted panel HH[b] (%)	Reached panel HH[c] (%)	Panel HH cooperation rate[d] (%)	Panel HH refusal rate[e] (%)	DBS consent rate[f] (%)	Median minutes per interview[g]
1	27	8	100	100	75	0	82	136.3
2	16	14	79	79	**45**	0	**57**	80.9
3	10	25	83	83	60	**33**	85	78.3
4	24	12	100	83	**45**	**28**	78	**48.3**

[a] HH stands for household.
[b] At least 99% of households need to be attempted for contact after 30 days in the field.
[c] At least 90% of panel households have to be contacted (i.e. reached) after 30 days in the field.
[d] At least 60% of all households are expected to cooperate after at least 70% of all households have been reached.
[e] Refusal rate is expected to be less than 16%.
[f] Dried blood spots (DBS) consent rate is expected to be 60% or higher.
[g] Median interview length is expected to be 60 min or longer.

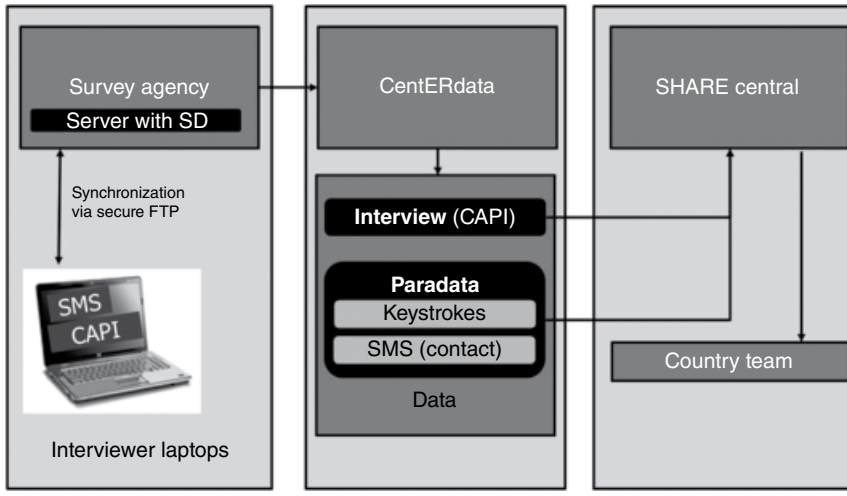


Figure 35.5 Data flow during ongoing fieldwork of SHARE.

tables use laptop IDs to refer to interviewers for anonymity purposes and to maintain data protection in all SHARE. These interviewer-level tables are emailed to each survey agency and country team with the request to take corrective actions for interviewers flagged on a given indicator.

Table 35.4 shows an example of an interviewer-level report used for wave 6 of SHARE in an undisclosed country (for the sake of clarity, not all indicators are displayed). Bolded rates indicate "underperformance" on the respective indicator. Underperformance is determined by not meeting predefined cut-off scores on the respective indicators (see Table 35.4). Interviewer with laptop (1) has satisfactory results. Interviewer (2), however, seems to have difficulty with obtaining respondents' consent to the collection of dried blood spots (DBS), a special survey request of wave 6. Note that while household cooperation rate also seems low (45%), this interviewer has not yet attempted to contact 21% of the assigned households and the refusal rate of reached households is 0%. Interviewer (3) seems to struggle with refusals: Almost one third of the reached households refuse to cooperate. Most concerning is interviewer (4) who seems to struggle with refusal rate and DBS rate and has an alarmingly short interview duration (median in wave 6 is about 75 minutes). The three flagged interviewers require different kinds of interventions. Interviewer (2) might require retraining on getting consent for DBS. Interviewer (3), depending on the causes of the high refusal rate, might either need to be retrained on recruiting respondents or might require assigning his or her households where there was a refusal to a different interviewer. Interviewer (4) requires special attention; the survey agency needs to

investigate the root cause(s) of the short interview duration and the quality of the conducted interviews through phone verification.

It is important to note that legally all corrective managerial actions are left to the fieldwork departments of the survey agencies. At the time of writing, no empirical investigation on the effectiveness of any performed interventions was completed. However, written feedback on explanations, actions taken, and their results is solicited from all countries during fieldwork. For example, in one situation, the high refusal rate for one of the interviewers was deemed to be the result of a difficult PSU that has inner urban multiunit dwellings. In another situation, a number of initially underperforming interviewers were flagged and, after questioning their work, either left the study voluntarily or were suspended because they were judged to be unqualified for a complex study such as SHARE. Other instances resulted in improvement in the DBS consent rate after a number of underperforming interviewers on this indicator were retrained.

Finally, after fieldwork, all survey agencies are asked to participate in an online survey about the agency's procedures regarding recruiting and managing interviewers, interviewer payment structures, and experiences with conducting the specific wave of SHARE. Ideas for improving SHARE are also solicited. Of the 18 countries (survey agencies) in wave 6, 15 (83%) reported that the interviewer-level statistics were useful for their fieldwork management. The three less satisfied countries found such statistics less useful for various reasons: One found them "too oppressive with only pointing out the negative […]," one agency reported that various interviewers share the same laptop so statistics on the laptop level were not strictly attributable to one-and-the-same interviewer, and finally one country complained about the information requested being an "overkill."

Several insights were gained from implementing the proactive interviewer-level monitoring in wave 6 of SHARE. These insights include recognizing the challenges of implementing an interviewer-level intervention when interviewers are not hired by the coordinating center (i.e. SHARE Central in Munich) and when they are part-time employees of the survey agency. First, any intervention requires a mediated communication through the local survey agency as these agencies are legally independent contractors to SHARE. Second, almost all survey agencies working for SHARE hire their interviewers on a self-employed, part-time basis and give the interviewers the freedom to manage their availability on the different projects they are assigned to. This affects the productivity of interviewers on SHARE as they tend to choose to spend more time on other (perhaps more profitable and less demanding) projects. Third, it is important to handle the initial reaction of the survey agencies to the tight and frequent management approach practiced by SHARE Central. This QC approach was initially seen as threatening to the agencies' autonomy. This challenge was, however, mitigated by creating a mindset of "shared problem-solving for better outcomes."

35.2.5.4 Summary

Overall the proactive interviewer-level monitoring approach implemented in wave 6 was successful and is a step in the right direction. One of the strengths of this approach is that the QC indicators used are not solely based on data provided by the interviewers. Some are generated from process data (such as time stamps) produced automatically by the interview software. However, several improvements are warranted. First, more human resources are needed at SHARE Central to prepare and explain to survey agencies the interviewer-level statistics and interventions needed. Second, a more systematic follow-up procedure regarding interviewer-level interventions implemented by the survey agency would enhance the effectiveness of such a proactive QC approach. Third, streamlined procedures and guidelines are needed to integrate interviewer-level statistics into the selection of cases for verification. In wave 6, most households are randomly chosen for verification, albeit late during the fieldwork. SHARE's experience is that this strategy delivers too little too late and lacks efficiency and timeliness. Finally, the QC plan for wave 7 will also identify top performing interviewers and reward them with badges after fieldwork. All these measures combined would make this interviewer-level approach more successful and effective at reducing interviewer error.

35.3 Conclusion

Interviewers can be a main source of survey error, increasing both the bias and the variance of survey estimates. The variation of interviewer error across sites in multinational, multiregional, and multicultural contexts adds another layer of complexity and compromises the comparability of findings. This chapter presents different case studies that used technological advances to generate real-time data for monitoring interviewers' work during the data collection phase. Utilizing real-time data, traditional routine QC procedures on a random subsample were augmented with targeted selection of cases or interviewers. Such targeted selection could increase the efficiency of the QC process by focusing resources on cases or interviewers that display unusual data patterns potentially linked to deviations from the study protocol. A list of recommended steps is summarized below for researchers and practitioners interested in implementing a similar approach in multinational, multiregional, and multicultural contexts:

- Begin with identifying a set of critical-to-quality indicators. These indicators need to be timely and are proxy measures of different types of survey error including nonresponse, measurement, and coverage. Examples of indicators are provided in the different case studies. Researchers and practitioners are encouraged to evaluate these indicators and choose ones that are relevant to the design of their study and its cultural context.

- Decide on a flagging rule for each indicator that can indicate deviations from the desired protocol.
- Develop an infrastructure that can pool the data underlying these indicators (interviewing data, sample management system data, keystroke data, and so on) into a central location. An important feature of such an infrastructure is the ability to process the pooled data, automate flagging rules, and display the outcomes for review.
- For multisite projects, all participating sites need to install integrated software tools that generate the data needed for the indicators (e.g. data collection software and sample management system). The SHARE case study provides a good example for such an integrated system.
- Develop an output display that is easy to visualize and identify troublesome cases. Such output needs to be sent to the project manager in each site on a routine basis for a timely review.
- The project manager in each site is encouraged to form a holistic picture before taking any action based on such a QC output. The project manager is encouraged to evaluate multiple indicators and conduct further investigations to identify a special cause variation that is driven by an interviewer. Such a holistic assessment is important to avoid penalizing interviewers and demoralizing them for reasons beyond their control. Thus, a feedback loop between the outcome of the assessment and the traditional QC procedures (verification and evaluation) is needed. The SNMHS case study provides a good example of such an integrated process.
- Each site is then required to report back any intervention taken on any interviewer. Such feedback must be collected using standardized documentation in a systematic and routine matter. Sites are also encouraged to take action not only on underperforming interviewers but also on top performing interviewers providing them with rewards.

It is important to highlight that an integrated data-driven approach requires added up-front costs, additional human resources, some usability testing, additional training, and strong communication between the coordinating body and the local sites. Whether the added cost is absorbed by the efficiency resulting from targeting specific troublesome cases is a central question that needs to be further investigated. It is also important to recognize that such an approach is less than optimal in its current form and requires improvements especially when implemented across sites in a multinational and multicultural context. Future usability improvements include a reduction in the number of data-driven quality indicators, harmonizing the indicators across sites, simplifying the visual display of the indicators, developing programming codes that utilize free software (or software that is more common across international sites), providing detailed guidelines on how to link the data-driven procedures to the routine QC procedures, and standardizing real-time documentation of

interviewer-level interventions across sites. Most importantly and before disseminating further such an approach, the association between the data-driven quality indicators and survey error and the effectiveness of the interventions tied to these quality indicators need to be empirically tested and proven.

References

1 Hansen, S.E., Benson, G., Bowers, A. et al. (2016). Survey quality. In: *Guidelines for Best Practice in Cross-Cultural Surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan http://www.ccsg.isr.umich.edu/ (accessed May 3 2017).

2 Lyberg, L. and Stukel, D.M. (2010). Quality assurance and quality control in cross-national comparative studies. In: *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (ed. J.A. Harkness, M. Braun, B. Edwards, et al.), 227–249. Hoboken, NJ: Wiley.

3 Lyberg, L. and Biemer, P.B. (2008). Quality assurance and quality control in surveys. In: *International Handbook of Survey Methodology* (ed. E.D. de Leeuw, J.J. Hox and D.A. Dillman), 421–441. New York: Psychology Press.

4 Mneimneh, Z.N., Tourangeau, R., Pennell, B.-E. et al. (2015). Cultural variations in the effect of interview privacy and the need for social conformity on reporting sensitive information. *Journal of Official Statistics* 31 (4): 673–697.

5 Blom, A. and Korbmacher, J. (2013). Measuring interviewer characteristics pertinent to social surveys: a conceptual framework. *Survey Methods: Insights from the Field*. http://surveyinsights.org/?p=817 (accessed 6 March 2018).

6 Davis, R.E., Couper, M.P., Janz, N.K. et al. (2010). Interviewer effects in public health surveys. *Health Education Research* 25 (1): 14–26.

7 Groves, R.M. (1989). *Survey Errors and Survey Costs*, Chapter 8, 357–404. New York: Wiley.

8 Groves, R.M., Fowler, F.J. Jr., Couper, M.P. et al. (2009). *Survey Methodology, 2*. New York: Wiley.

9 West, B.T. and Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly* 75 (5): 1004–1026.

10 Groves, R.M. and Magilavy, L.J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly* 50: 251–266.

11 Buellens, K. and Loosveldt, G. (2016). Interviewer effects in the European social survey. *Survey Research Methods* 10 (2): 103–118.

12 Davis, P. and Scott, A. (1995). The effect of interviewer variance on domain comparisons. *Survey Methodology* 21 (2): 99–106.

13 O'Muircheartaigh, C. and Campanelli, P. (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society, Series A* 162 (3): 437–446.

14 Schnell, R. and Kreuter, F. (2005). Separating interviewer and sampling point effects. *Journal of Official Statistics* 21 (3): 389–410.

15 Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association* 57: 92–115.

16 Blom, A.G., de Leeuw, E.D., and Hox, J.J. (2011). Interviewer effects on nonresponse in the European Social Survey. *Journal of Official Statistics* 27 (2): 359–377.

17 Loosveldt, G. and Beullens, K. (2013). How long will it take? An analysis of interview length in the fifth round of the European Social Survey. *Survey Research Methods* 7 (2): 69–78.

18 Stoop, I., Billiet, J., Koch, A., and Fitzgerald, R. (2010). *Improving Survey Response: Lessons Learned from the European Social Survey*. Hoboken, NJ: Wiley.

19 Mneimneh, Z.N. (2012). Interview privacy and social conformity effects on socially desirable reporting behavior: importance of cultural, individual, question design and implementation factors. Unpublished dissertation. University of Michigan.

20 Crespi, L.P. (1945). The cheater problem in polling. *Public Opinion Quarterly* 9 (40): 431–445.

21 American Association for Public Opinion Research (2003). Interviewer falsification in survey research: current best methods for prevention, detection and repair of its effects. https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/falsification.pdf (accessed 7 March 2018).

22 Couper, M.P. (1998). Measuring survey quality in a CASIC environment. *Proceedings of the Survey Research Methods Section, American Statistical Association* 48: 743–772.

23 Kreuter, F. ed. (2013). *Improving Surveys with Paradata: Analytic Uses of Process Information*. Hoboken, NJ: Wiley.

24 Durrant, G.B., D'Arrigo, J., and Steele, F. (2011). Using paradata to predict best times of contact, conditioning on household and interviewer influences. *Journal of the Royal Statistical Society, Series A* 174 (4): 1029–1049.

25 Groves, R.M. and Heeringa, S.G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, Series A* 169: 439–457.

26 Kreuter, F., Couper, M.P., and Lyberg, L.E. (2010). The use of paradata to monitor and manage survey statistical data collection. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 282–296. Vancouver, BC (31 July–5 August).

27 Luiten, A. and Schouten, B. (2013). Tailored fieldwork design to increase representative household survey response: an experiment in the Survey of Consumer Satisfaction. *Journal of the Royal Statistical Society, Series A* 176 (2): 169–189.

28 Bushery, J.M., Reichert, J.W., Albright, K.A., and Rossiter, J.C. (1999). Using date and time stamps to detect interviewer falsification. *Proceedings of the*

Survey Research Methods Section, *American Statistical Association*, 316–320. Baltimore, MA (8–12 August).

29 Hood, C. and Bushery, M. (1997). Getting more bang from the reinterviewer buck: identifying "at risk" interviewers. *Proceedings of the Survey Research Methods Section*, *American Statistical Association*, 820–824. Anaheim, CA (10–14 August).

30 Josten, M. and Trappmann, M. (2016). Interviewer effects on a network-size filter question. *Journal of Official Statistics* 32 (2): 349–373.

31 Kosyakova, Y., Skopek, J., and Eckman, S. (2015). Do interviewers juggle filter questions? Evidence from a multilevel approach. *International Journal of Public Opinion Research* 27: 417–431.

32 Matschinger, H., Bernert, S., and Angermeyer, M.C. (2005). An analysis of interviewer effects on screening questions in a computer assisted personal mental health interview. *Journal of Official Statistics* 21: 657–674.

33 Murphy, J., Baxter, R., Eyerman, J., Cunningham, D., and Kennet, J. (2004). A system for detecting interviewer falsification. Paper presented at the 59th Annual Conference, American Association for Public Opinion Research. Phoenix, AZ (13–16 May 2004).

34 Thissen, M.R. and Myers, S.K. (2015). Systems and processes for detecting interviewer falsification and assuring data collection quality. *Statistical Journal of the IAOS* 32 (3): 339–347.

35 Bredl, S., Storfinger, N., and Menold, N. (2011). A literature review of methods to detect fabricated survey data. *Discussion Papers from Justus Liebig University Giessen*, Center for International Development and Environmental Research (ZEU).

36 Kessler, R.C. and Üstün, T.B. (2004). The World Mental Health (WMH) survey initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research* 13: 93–121.

37 WHO (2017). The World Health Organization World Mental Health Composite International Diagnostic Interview (WHO WMH-CIDI). http://www.hcp.med.harvard.edu/wmhcidi/ (accessed 7 March 2018).

38 Harkness, J.A., Villar, A., and Edwards, B. (2010). Translation, adaptation and design. In: *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (ed. J.A. Harkness, M. Braun, B. Edwards, et al.), 117–140. Hoboken, NJ: Wiley.

39 SQP (2017). Survey quality predictor. http://sqp.upf.edu/ (accessed 7 March 2018).

40 Groves, R. and Couper, M. (1998). *Household Survey Nonresponse*. New York: Wiley.

41 cApStAn (2017). cApStAn. http://www.capstan.be/ (accessed 7 March 2018).

42 Kish, L. (1994). Multipopulation survey design. *International Statistical Review* 62: 167–186.

43 Groves, R. and Lyberg, L. (2010). Total survey error: past, present and future. *Public Opinion Quarterly* 74 (5): 849–879.
44 Malter, F. (2014). Fieldwork monitoring in the Survey of Health, Ageing and Retirement in Europe (SHARE). *Survey Methods: Insights from the Field*. http://surveyinsights.org/?p=1974 (accessed 7 March 2018).
45 SHARE (2014). SHARE compliance profiles – wave 5. http://www.share-project.org/fileadmin/pdf_documentation/SHARE_Wave5_ComplianceProfiles_v11.pdf (accessed 7 March 2018).