Quality Assurance Protocols for Health, Nutrition and Demographic Surveys

Laxmikant Dwivedi

Table of Content

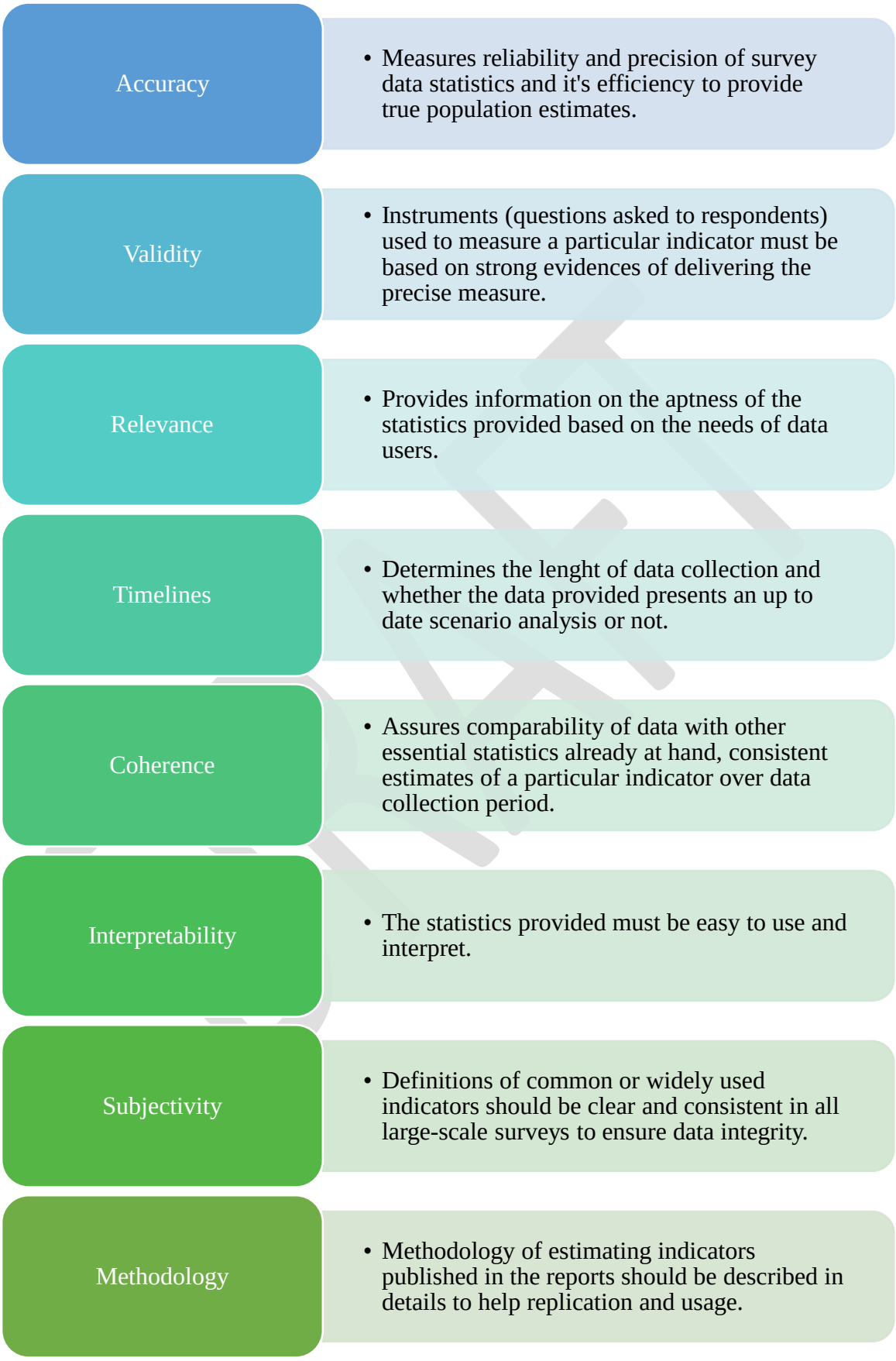
List of Figures

Chapter 1: Introduction- Importance of data quality checks in large-scale surveys

Large-scale sample surveys collecting data on public health indicators play a pivotal role in monitoring and evaluation of ongoing public health policies and programmes, effectiveness of resource allocation and situational analysis of various public health indicators. Therefore, it is only fundamental that large-scale surveys invest resources in training and monitoring data collection process and cleaning of data to ensure good quality and reliable data collected from the field. The data that is made publicly available goes through a detailed process of micro-management and constant supervision with the aim of improving its quality. In this report, various data quality assurance measures taken in large-scale surveys and some future measures that can be taken to improve data quality and provide even more reliable and valid data that accurately captures current public health scenario are discussed. For ease of comprehension, this report is divided by survey timeline. At the beginning of the survey, in the planning phase, measures to ensure that the data collected from the survey is accurate and devoid of misreporting are taken. On field, during the ongoing process of data collection, several surveys take a variety of steps to guarantee that the data that is being collected by the investigators are correct and true to the existing notions based on theoretical underpinnings. Deviation from the expected norms is checked and registered by the experts. After the entire fieldwork is over, the collected data is cleaned and processed to make it presentable and usable by academicians, researchers, policy makers, national and international funding agencies and public administrators. Good quality data delivery by large-scale surveys requires constant dedication, supervision and monitoring along with robust sampling design and statistical decision making. Data quality of a large-scale survey may be compromised due to sampling and non-sampling errors. Sampling errors are those arising due to the sampling frame and design selected for the survey. Non-sampling errors include non-response or

misreported responses by respondents of the survey, bias due to incorrect reporting by the investigators, faulty questionnaire designs, recall bias. In the following sections, the focus will be on ways adapted by various surveys to deal with these errors. We discuss the need for maintaining data quality protocols in all the steps of a survey and provide examples from various large-scale surveys.

Figure 1. Pillars of high-quality survey data



Examples of best practices are drawn from a few of the popular data sources. In the following section the data sources used for examples are briefly described.

1. **National Family Health Survey (NFHS)**

   Demographic Health Surveys (DHS) are conducted in several countries to collect nationally representative data on population, health and nutrition. The major areas of data collection are demographic characteristics, education and wealth status, adult health and nutrition, sexual and reproductive health indicators, family planning and unmet needs, fertility and related choices, child health and nutrition, infant and child mortality, maternal health and mortality, women empowerment, domestic violence, HIV prevalence and its related knowledge, attitudes and behavior, etc. In India, the DHS is named as the National Family Health Survey (NFHS). The first NFHS was conducted in 1992-93. These surveys are conducted under the stewardship of Ministry of Health and Family Welfare (MoHFW) and the International Institute for Population Sciences (IIPS), Mumbai has been the nodal agency.

2. **Comprehensive National Nutrition Survey (CNNS)**

   The nationally representative Comprehensive National Nutrition Survey (CNNS), 2016-18 was implemented in India by the Ministry of Health and Family Welfare, Government of India and supported by United Nations Children's Fund (UNICEF). The central implementing agency, the Population Council and four survey agencies (KANTAR Public, Gfk Mode Pvt. Ltd, SIGMA Research and Consulting Pvt. Ltd and the Indian Institute of Health Management Research, Jaipur) were involved in data collection in 30 states and Union Territories of India. Anthropometric measures of 112,316 children and adolescents were collected. Biological samples (blood, urine, stool) were collected from 51,029 children and adolescents.

3. Sample Registration System (SRS)

SRS is one of the most reliable sources of data for regular vital statistics of India. The main objective of SRS is to furnish regular and reliable estimates of birth rate, death rate, infant mortality rate, total fertility rate at the natural division level for the rural areas and at the state-level for the urban areas. Natural divisions are defined as the NSS classified group of contiguous districts with diverse topographical characteristics. The survey also provides periodical data on maternal deaths in India and causes of deaths through post-death verbal autopsy method of data collection.

4. Longitudinal Ageing Study in India (LASI)

Longitudinal Ageing Study in India (LASI) is a panel study whose main waves (I and II) is decided to take place between 2016-2021. Then follow-up interviews will be conducted biennially until 2040. The project was approved in 2013 under the sponsorship of MoHFW (NCD division, GOI), National Institute on Ageing, National Institute of Health (USA) and UNFPA. IIPS is the nodal agency for the survey.

5. National Sample Surveys (NSS)

The Directorate of National Sample Survey came into existence in 1950 under the Ministry of Statistics & Programme Implementation (MOSPI). However, the National Sample Survey Organization (NSSO) was created under a government set-up in 1970. In 2006, the Governing council of NSSO was dissolved and all the responsibilities was handed over to the National Statistical Commission (NSC). NSSO is now headed by the Director General of Survey who holds the entire responsibility of the mandate of the NSSO.

6. Global Youth Tobacco Survey (GYTS)

GYTS is a school-based survey with the central aim of enhancing capacity of countries for surveillance, monitoring and guidance to implement and evaluate tobacco prevention and control programmes among youth aged 13-15 years studying in grades 8-10. The first round was at state-level in 2000-2005 and the next two rounds were at the national level in 2006 and 2009. International Institute for Population Sciences (IIPS), Mumbai has been appointed as the nodal agency for the ongoing round by MoHFW. GYTS provides essential information on knowledge, attitude and use of smoke and smokeless tobacco among the youth, exposure to media coverage on tobacco use, health impact, secondhand smoke, tobacco related school curriculum, cessation of use of tobacco.

7. Indian Human Development Survey (IHDS)

Till date there has been two rounds, IHDS I (2004-05) and II (2011-12). It is a collaborative research programme between the research team from the National Council of Applied Economic Research and University of Maryland. It is the first panel survey in India that follows up with the respondents and households over the panel period. The main aim of IHDS is to capture changes in society at large and their influence on households, life cycle changes and transition in human lives.

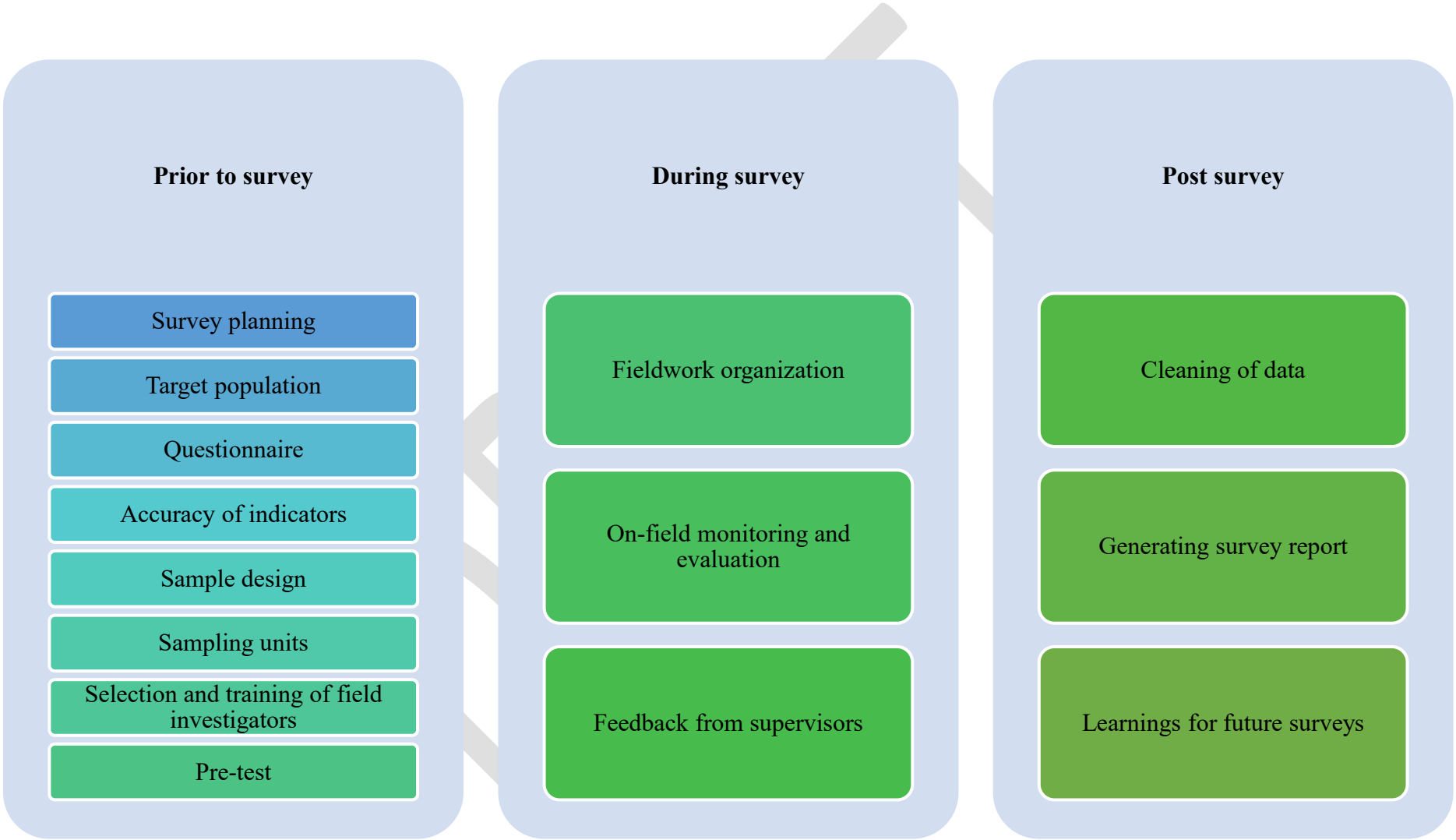
8. National Integrated Biological and Behavioral Surveillance (IBBS)

National AIDS Control Organization (NACO) is in charge on the National IBBS which is one of the largest bio-behavioral study of its kind. The central aim of the survey is to provide HIV prevalence and identify risk behavior among the high-risk groups (HRG), evaluate the ongoing National AIDS Control Programme (NACP) and support planning and prioritization of programme efforts at the district, state and national levels. The survey also aims to measure the change in HIV related risk behaviors among the HRG between the baseline and endline of NACP IV (2012-2017). IBBS identifies Female Sex Workers (FSW), Men who have sex with men (MSM), Injecting Drug Users (IDUs), Transgenders (TG), and Migrants and Currently Married Women in high outmigration districts as HRG.

9. Understanding the lives of adolescents and young adults (UDAYA)

UDAYA is carried out in Bihar and Uttar Pradesh to understand the levels, patterns and trends in the situation of younger (10-14 years old) and older (15-19 years old) adolescents and assess the transitions in their life. The survey is conducted by the Population Council on married and unmarried adolescents of the age group 10-19 years. It is funded by funded by the Bill and Melinda Gates Foundation and the David and Lucile Packard Foundation.

Figure 2. Conceptual framework of data quality checks in large-scale surveys



Chapter 2. Pre-survey data quality checks

Prior to any large-scale survey, there are several data quality control measures that are integrated to ensure smooth data collection and minimization of sampling and non-sampling errors. Most of these decisions are based on expertise of the survey coordinators and is done in the planning phase itself. It is of utmost importance to select the target population for the sample survey and design a questionnaire that can be appropriate for the sample. Special care and attention have to be given in sequencing, placement and wording used in the questionnaire. Since most large-scale surveys have a substantial coverage in various states of India, consistency in meaning of questions after translation in vernacular language has to be maintained. Some surveys attempt to capture sensitive indicators like domestic violence, sexual health. It is required to pre-decide the accuracy level of such sensitive measures that are acceptable to provide statistical estimates. To ensure high quality of data, sample design, error measurements and bias considerations are required to be estimated through statistically sound tools and methods. The total sample size for the survey, divided into broad categories, like number of households to be interviewed, number of individuals by age-groups and gender to whom questions will be administered are some of the sampling unit decisions that are to be made during the survey planning phase. The survey organizers then interview and employ experienced investigation organization who in turn provide efficient investigators who are deemed to be capable in data collection. Most large-scale surveys undertake rigorous training of these investigators to provide clear instructions on how each question is to be administered, how respondents should be encouraged to provide correct responses, environment to be maintained during on-field interviews, methods to obtain certain biometric measures, etc. After considering robust sampling techniques, reliable and clear questions and incorporating all the issues that experts apprehend to arise during data collection phase, the questionnaire has to be pretested on a small sample. Pretesting or a pilot survey helps in

identifying red flags in the questionnaire and provides the chance for a final revision before the actual fieldwork begins.

1. **Survey planning**

At the start of planning of global surveys involving multiple countries, such as Demographic Health Survey (DHS), (SAGE), Global Youth Tobacco Survey (GYTS) implementing and nodal agencies are selected at various national levels. These agencies are ideally Ministries of health/ education or academic institutions with experience of conducing such large-scale surveys. Other factors like availability, timeline management capacities, collaborations with Government Agencies are also considered.

A capable Research Coordinator is then selected to manage the entire survey implementation encompassing data collection, cleaning, analysis, report publication and data dissemination. Most surveys provide a detailed criteria of selection of Research Coordinator to ensure integrity and credibility of the data collected by the survey. The criterion takes into account the experience and skills of the Research Coordinator, availability to participate in training workshops, collaborations with Government Agencies, sound knowledge of sampling techniques and epidemiology (in case of health surveys) and updated with current scenario of the survey objectives.

Before a survey is conducted a team of experts come together to brainstorm the objectives of the survey and zero down on the essential key findings the survey is expected to deliver. It is essential to have a clear goal and expectation to generate precise questions and measures for data quality maintenance.

Examples:

NFHS

NFHS 4 covers an array of topics such as health, family planning, family welfare, reproductive and sexual health, child and maternal morbidity and mortality. With every consecutive survey rounds, there is an increase in the sample and coverage of NFHS. While this reduces the sampling errors, there is always a risk of increase in the non-sampling errors. Hence, during the survey planning stages, deciding the sample size has to be balanced on the demands of analysis, accuracy and the capacity of the implementing organization and fund constraints.

In the last two NFHS rounds - 4, 5 the content and number of questions canvassed has increased substantially to capture various new initiative and policies by Government. Simultaneously, from being a state level survey, it shifted to provide district level data which increased the sample size considerably. Field operations was carried out in two phases, larger states were divided into 2-3 parts and survey was implemented in a cluster of 5-6 adjoining districts.

CNNS

In the planning phase of the survey, it is a good practice to document data gaps of existing surveys. CNNS provides data on nutrition and health of children and adolescents and was published around the same time as the NFHS 4. CNNS identified data gaps in the following sectors:

1. Micronutrient deficiencies across younger age groups
2. Causes of anaemia among children
3. Non-availability of nutrition indicators in 5-19 age groups
4. Non-communicable diseases among children below 14 years

5. Lipid profile of school going children to assess risk of cardiovascular diseases
6. Indicators of chronic kidney diseases among school-aged children
7. Measures of nutrition status among children other that height/ weight for age z-scores, such as waist circumference, skinfold thickness, grip strength, physical fitness.

Documenting such data gaps can be beneficial in providing more accurate and useful information by a survey. It can also be an effective tool to narrow down questions to retrieve detailed data on specific objectives of the survey.

SRS

Prior to the start of survey and continuous enumeration, SRS conducts a baseline survey to prepare a notional map of the enumeration area, house listing, filling out the household schedule and listing of women in reproductive ages in the enumeration area.

LASI

The objective of the survey is to cover domains on health, economic conditions and social connection of adults and older population in India. For this purpose, it is designed as a panel study that will follow a cohort of adult population from the pre-retirement age of 45 years and above to older ages of 60 years and above. It is one of the largest longitudinal ageing studies. All states are divided into smaller regions but union territories were not divided. The survey is to be carried out in 2 phases for maximum utilization of resources and operational efficiency.

NSS

NSS is a body that conducts different surveys and thus, it has various divisions under it to perform different roles and have varying responsibilities. There are several active divisions under NSSO in charge of various activities:

1. Coordination & Publication Division (CPD), New Delhi: This division is the nodal division for coordination with Ministries, State Government, other statistical agencies. This division is responsible for the overall monitoring of Capacity Development (CD) Scheme, a Central Sector Scheme along with Support for Statistical Strengthening (SSS), a Sub-Scheme of CD Scheme. The goal of this scheme is development of infrastructure, improving the Statistical Capacity and Infrastructure of the State Statistical System as well as mobilizing manpower to provide credible, reliable and timely official statistics. This schemes also encourage usage of these statistics at the State/District and Block Levels. This division also supervises the administrative work of Indian Statistical Institute (ISI).
2. Data Informatics and Innovation Division (DIID): This division is mandated with Electric Data Processing (EPD) including data processing, data management, software development, processing, validation and tabulation of sample survey data collected. It also archives and disseminates data as per the international standards.
3. Data Quality Assurance Division (DQAD): Erstwhile Data Processing Division (DPD), is primarily mandated for ensuring data quality of the socio-economic surveys. It has five centers at five different cities with its Headquarters in Kolkata. Currently it is in charge of the 78th Round NSS Survey: Domestic Tourism Expenditure and Multi Indicator Survey, Periodic Labor Force Surveys (PLFS), Annual Surveys of Unincorporated Sector Enterprise (ASUSE), Time Use Surveys (TUS), Household Consumption Expenditure Surveys (HCES), Annual Survey of Service Sector Enterprise (ASSSE).
4. Economic Statistics Division (ESD): The ESD is responsible for calculating the Index of Industrial Production (IIP) using data from 14 source agencies in various Ministries/Departments or their attached/ subordinate offices. IIP is a composite

indicator to measure the short-term changes in the volume of production of a basket of industrial products during a given period with respect to that in a chosen base period. ESD is also responsible for conducting the periodic Economic Census in India to develop a database for all non-farm economic establishments. This division compiles Energy statistics annually and provides updated data, provides technical guidance to Ministries to develop Index of Service Production (ISP) at sector/ sub-sector levels, develops classifications like National Industrial Classification (NIC) and National Product Classification (NPC) based on revised guidelines of International Standard Industrial Classification (ISIC).

5. Field Operations Division (FOD), New Delhi: This is the biggest division of NSS with 6 zonal offices, 52 regional offices and 117 sub-regional offices. The main responsibility of this division is to coordinate with various field agencies and carry out field work of all the socio-economic and other surveys under NSS.
6. National Accounts Division (NAD): NAD publishes the annual National Accounts Statistics reports containing Gross Domestic Product (GDP), National Income, Government/Private Final Consumption Expenditure, Capital Formation and Savings along with details of transactions of institutional sectors. They are also responsible for the production of Supply-Use Tables (SUT) and Input-Output Transaction Tables (IOTT). This division provides technical guidance and support to the State/UT Directorates of Economics & Statistics (DESs) on compilation and release of State Accounts.
7. Price Statistics Division (PSD): Compiles and releases all India and state-wise Consumer Price Index (CPI), provides technical guidance to states and Ministries for price indices.

8. Social Statistics Division (SSD): SSD deals with development of social, environment and multi-domain statistics. It also has a Sustainable Development Goal (SDG) unit that provides status of sub-national regions in regard to the SDGs.
9. Survey Coordination Division (SCD): Monitors allocation of Grants-in-Aid to Northeastern states of India. This division is also responsible for publication of the bi-annual journal of NSS "Sarvekshana". It also organizes seminars to discuss research carried out using various NSS datasets.
10. Survey Design and Research Division (SDRD): This department is responsible for sampling design, recognizing concepts and definitions, drawing up survey schedules, framing instruction manual for field work, finalization of sample list, scrutiny and validation instructions, checking of multipliers and tabulation plans, finalizing estimation procedures and survey results, preparation of key report, revision and improvement of survey methodology and techniques, and secretarial assistance to the working group on NSS.
11. Training Division: Under the training division, National Statistical Systems Training Academy (NSSTA) has been established as a premier institute to foster and develop human resource in collection and application of official statistics. This division promotes various activities to boost the use of NSO statistics and monitors the implementation of Grant-in-Aid component of the Capacity Development Scheme. The basic objective of NSS has been to collect nation-wide data on various socio-economic subjects to facilitate policy formulation, programme implementation and evaluation, research and, economic and administrative decision making. NSSO generally has a ten-year cycle in which they collect data on different topics:
    (i) Consumer Expenditure and Employment & Unemployment- twice
    (ii) Social Consumption (health, education etc.) (generally 6-month survey)- twice

(iii) Un-organised Manufacturing- twice

(iv) Services Sector- twice

(v) Land & Livestock holdings and Debt & Investment- once

(vi) Open Round - Two years (For other surveys on demand)

IBBS

NACO is the main agency under which IBBS is carried out. At the national level, a Technical Advisory Group (TAG) comprise a senior staff from NACO, regional public health institutes and development partners. The TAG supervises and monitors policy needs, technical strategies and smooth implementation of the survey. A National Working Group (NWG) consisting of members from different disciplines including NACO was responsible for design and development of methodologies, and guidelines for survey implementation. Eight regional institutes were identified as nodal agencies to implement IBBS. Each of the nodal points were allocated 2 to 5 states.

2. Target population

Many surveys have a pre-decided target population which is to be sampled. In some surveys, age of respondents maybe a selection criteria. Ideally, it is best advised, that the survey addresses data gaps and not collect repeated information that is already collected by other surveys. For this purpose, before the start of the survey, a detailed document on previous surveys, data present and gaps in information must be prepared. Target population is usually decided upon by confirming the objectives of the survey and the policy issues the survey aims to address. Selection of target population depends on the varied objectives of the survey and defines the sampling frame of the survey. A few surveys also provide incentives to ensure that the sampled respondents stay motivated to respond to all the questions in the survey.

Examples:

NFHS

The major aim of NFHS is to provide reliable maternal and child health indicators. In the last two rounds, NFHS has also provided data on health, socio-economic aspects of males. For this purpose, the survey gives emphasis on collecting data from 15–49-year-old women and 15-54 year old men.

SRS

The continuous enumeration approach requires the enumerator to maintain the birth and death record for the sampled households in their respective areas. The enumerator is expected to record all births and deaths pertaining to (i) usual residents inside the sample unit; (ii) usual residents outside the sample unit; (iii) in-migrants present; (iv) in-migrants absent; and (v) visitors inside the sample unit (however, this is not included while calculating estimates).

LASI

LASI aims to capture the transition of adults to older ages and to provide data to understand life course events and their long-term effects on the elderly population in India. To meet this goal all men and women aged 45 years and above will be interviewed in the selected households. Spouses of the respondents will also be interviewed irrespective of their ages. LASI has chosen the age 45 to make the survey comparable to Health Retirement Study (HRS) surveys in Asia and to capture transition in the consumption-expenditure behavior from pre-retirement to post retirement.

GYTS

The target population for GYTS is 13- to 15-year-old students. The objective of the survey is to gather information on tobacco use among school going adolescents and produce cross-sectional national level estimates. For this purpose, they withdraw sample from school going population aged 13-15 years.

IBBS

Considering the sensitive nature of the survey, ethical statements are of utmost importance in almost all large-scale surveys. Through this target population can be reassured of confidentiality and privacy of the data they provide and the results of the tests they consent to during the survey. For IBBS, INR 200 was provided to the respondents to compensate for their time (including transportation). Financial incentives might motivate the respondents to be more engaged in the survey. IBBS also adopted the “unlinked anonymous” approach to ensure confidentiality of the respondents. Owing to the sensitivity of the survey, IBBS has also adopted the “community first” approach as suggested by NACP. In order to engage communities and maximize community involvement in the survey, Community Advisory

Boards (CAB) and Community Monitoring Boards (CMB) were formed. This aided in collecting good quality data.

3. Questionnaire

Questionnaires of global surveys often have set of pre-fixed questions that are repeated for each round of the survey. However, many questions are adapted for country-specific use to capture national level scenarios.

Surveys usually provide specific guidelines to adhere to while adapting or translating questions for country-specific use. Surveys also consider the negative impact on data quality by increasing the number of questions. Hence, many surveys have an upper limit of total number of questions to be included in order to avoid respondent fatigue resulting in an increase in missing response. Based on the objectives of a particular survey, they may have some special supporting questionnaires too.

Another important issue to keep in mind while designing questionnaires is the sub-sections. Usually, a survey attempts to capture a particular aspect through a battery of questions. Questionnaires should be separated based on various aspects of the survey objectives.

Examples:

NFHS

To ensure data quality, the recent surveys have developed a nested design with a modular approach. Here, some of domains were kept in the state module, whereas some were in the district module. District module was a sub-set of the state module. Domains like husband's background, women's work, empowerment status were included in the state module as it was not the priority of the survey to provide district level estimates for these domains.

CNNS

For ease of data collection, survey questions were constructed both in English and in the state vernacular. Two questionnaires were monitored: household and individual. To avoid

unnecessary bulk of ineligible questions, the individual questionnaire was divided by age groups: 0-4 years, 5-9 years, 10-14 years and 15-19 years. This division by age group helped in separating out questions that were solely apt for a particular age group. For example, Dietary intake: breastfeeding, complementary feeding and dietary diversity for 0-4 year old preschool children and only dietary diversity for 5-19 year old school age children and adolescents.

LASI

LASI has three questionnaires: household schedule, individual schedule and community schedule. Community schedule is to collect information on the community the respondent resides in from eminent social and political figures. This schedule will provide data on population, socio-economic characteristics, infrastructure facilities, social and welfare programmes in the selected communities.

NSS

In order to collect accurate data free of respondent driven bias, NSS has covers an array of topics using different surveys and short questionnaires. This helps in providing accurate estimates of several questions on health, dietary pattern, expenditure patterns and economy.

GYTS

Many surveys pre-decide on the upper limit of the number of questions to be included. Like GYTS suggests a maximum of 75 questions for the country questionnaire to ensure data quality. Since GYTS includes school going adolescents it has a seven-question long School Policy questionnaire to capture the policies in place at different schools to tackle use of tobacco among students along with the regular questionnaires.

IHDS

IHDS I had three different questionnaires: the household questionnaire, the woman schedule and the institutional modules. Under the institutional modules there were village surveys, medical surveys and primary school surveys. Along with these existing questionnaire IHDS II added a youth module to capture education, employment, skill set, social network, marriage, risk behavior among those who were 8-11 years of age during IHDS I.

UDAYA

In UDAYA, there are two basic questionnaires: a household-level questionnaire, where the household head or an adult's responses are taken and an individual-level questionnaire used to interview the adolescents identified from these households. The individual-level questionnaire has four variants administered to unmarried boys aged 10-14 and 15-19 and, unmarried girls aged 10-14 and unmarried and married girls aged 15—19. This helps in narrowing down of questions in each of these variants and bolster the investigators capacity to monitor questions with more clarity.

4. Accuracy of indicators

In order to improve accuracy of indicators, survey instruments and collection techniques are constantly revised. Along with improvement in survey collection tools over a period of time, there has been significant changes in data collection and registration of life events. For example, in earlier DHS rounds, birth displacement and omission of births was a data quality concern. However, with time, birth registration has improved, and birth date of children can be easily recorded from a government issued document.

Examples:

NFHS

NFHS is transparent in showing the standard errors and confidence intervals of various essential indicators. This is usually included in the Appendix of the report. Skip patterns, filters, eligibility for sections are handled by CAPI. Introduction of CAPI has been a paradigm shift in the data collection process that not only ensures more systematic data collection but also has decreased the time taken to conduct interviews and provides data in exportable formats that can be easily used in any statistical software during the field work. It automatically handles skip patterns, filters, eligibility for sections. Protocols for collection Biomarker data has been developed as per international standards and is comparable with other DHS surveys. High quality and self-calibrated equipment are used for collection of data on biomarkers. The instruments were standardized periodically to maintain their accuracy.

CNNS

To maintain accuracy of various anthropometric indicators, highest quality of instruments was used.

1. Height/length-Three-piece wooden board
2. Weight-Digital SECA scale

3. Mid-upper arm circumference (MUAC)- MUAC tape
4. Triceps skinfold thickness (TSFT), Subscapular skinfold thickness (SSFT)- Holtain Skinfold Calipers
5. Waist circumference (WC)- Fiberglass tape

All the measures were taken twice but weight was measured once, and hand grip strength was recorded thrice.

In case of biological samples, fasting samples were collected. Two out of three children among 5-19 years were selected through systematic random sampling and all children between age 1-4 years were selected based on the pilot test.

SRS

The sample size of the SRS was previously based on the reliability of birth rate. From 2004 onwards, it is based on the reliability of infant mortality rate. The enumerator triangulates information on vital events from various resources and also visits all households once in each quarter (for rural areas) and once in a month (for urban areas) to ensure that all events are recorded. They are also directed to maintain and update a list of all women along with their pregnancy status to keep track of all births occurring in their designated sample areas.

LASI

To maintain good quality of data the survey uses several technologies- data will be collected using CAPI which has its advantages and will smooth out the data collection process, the survey includes IT based technologies like use of geographic information system (GPS) and barcodes for matching and anonymizing data. LASI has divided the biomarker data into 4 categories: *functional health markers* (Blood Pressure and Pulse rate (CVD), Lung Function Test (OAD/Respiratory diseases), Vision Test :Near and Distance visual acuity); *DBS based*

markers (C-reactive Protein (CRP) (CVD), Epstein Bar virus/ Cytomegalovirus (EBV/CMV) (Immunity), Glycosylated Hemoglobin (HbA1c) (Diabetes), Hemoglobin (Hb) (Anaemia), Cystatin C (Kidney disease), Vitamin D (Bone disease)); *anthropometric measures* (height, weight (BMI), waist and hip circumference (WHR)); *performance-based markers* (grip strength, timed walk (frailty), balance test (cerebellar function)).

NSS

In most large-scale surveys validation lies at the heart of the data quality protocols. These validations check and verification of consistency is usually done after data entry for one "lot", i.e., schedules of one first stage unit (FSU which is like a PSU in case of NFHS), is complete. Data validation in NSS is done using a set of survey-based instructions and checks that are fed to the computer in the form of in-house developed and customized programmes. There are three phases of data validation that are usually carried out in an NSS round:

(i) Phase 1: Content check is done by Computer Scrutiny Programme that generates error lists based on inconsistencies in data and thereby corrections are made using instructions and ancillary knowledge.
(ii) Phase 2: Coverage check for each FSU and SSU along with duplication checks.
(iii) Phase 3: Howler check for identifying outliers (abnormally high or low values) by range checks.

To contain error in data collection, CAPI is used in all the surveys.

IBBS

IBBS collected blood samples of respondents using the finger prick method or dried blood spot (DBS) method. The survey followed the "tow test protocol" for HIV testing where all blood samples were tested for HIV and only those that were reactive in the first stage were

subjected to the additional second stage. The specimens that were reactive in both the tests were labeled as “positive”. The testing was done using validated enzyme-linked immunosorbent assay (ELISA) kits in the laboratories.

5. Sample design

A robust sampling design is the backbone of any large-scale survey. Sampling design is based on prior knowledge of indicators and desired consistency of estimates. Most large-scale surveys are unable to follow a simple random sampling owing to the vast data that needs to be collected to provide regional level estimates. For this, a design effect needs to be estimated that accounts for the increase in variance of an estimator due to the application of other sampling techniques like multistage sampling against simple random sampling which has the least variance. In order to provide appropriate and representative estimates, weighting is a major part of any sample-based survey. These weights are generated by taking the survey sample design at each of the sampling stages, like cluster, household, individuals. It is usually based on the probability of selection of a sampling unit.

Examples:

NFHS

For most of the large-scale sample surveys like NFHS, simple random sampling is not feasible. In a sample such as NFHS, sample size is inflated by using the design effect accounting for the complex structure and clusters. Design effect adjusts for the loss in efficiency due to cluster sample as compared to simple random sample. It is defined as the ratio between the standard error of an estimate using the survey sample design and the standard error that would have resulted if simple random sampling was done. In NFHS, 15% of the households formed the sub-sample for state module. To maintain randomness of the sub-sample, NFHS administered the state module in every alternate house in 30% of the enumeration area of the selected clusters.

CNNS

It has a multi-stage sampling design to select households and individuals between the age 0-19years across all the regions of India. In case of the rural sample, Primary Sampling Units (PSUs) were villages listed from Census, 2011. In the first stage, probability proportional to size (PPS) sampling was employed to select the PSUs. Stratified sampling (explicit and implicit stratification) was adopted in the first stage to ensure representation of various socioeconomic groups (similar to NFHS sampling design). Larger PSUs (300 households or more) were further subdivided, and two segments were randomly selected. In rural areas, villages with less than 10 households were removed from the sampling frame and villages with less than 150 households were linked with neighbouring villages and were sampled as linked PSUs to have a minimum of 150 households. In the next stage, systematic random sampling of households within each PSU was adopted.

For the urban sample, PSUs were wards listed from Census, 2011 stratified by geographical regions. In the first stage, urban wards selected from strata using PPS. Each ward consists of Census Enumeration Blocks (CEB) consisting of 100-150 households. Smaller CEBs were linked to obtain merged blocks with a minimum of 150 households. In the second stage, one CEB was selected from each of the selected wards. In the third stage, households were randomly selected from the selected CEBs. Proper sampling design ensures representativeness of the estimates derived after data collection.

SRS

SRS follows a uni-stage stratified simple random sample without replacement except in larger villages (stratum II villages with population size more than 2000) where two stage stratification is done. In bigger states, the NSS natural divisions form the first geographical strata. Villages with population less than 200 are excluded from the sampling frame such that

they did not exceed 2 percent of the total population of the state. The villages were implicitly stratified into 3 equal sub-strata using their population sizes and ordered by their female literacy rates. The sample villages within each sub-stratum were selected randomly. Villages belonging to stratum II were subdivided into 2 or more segments such that none of the segments were divided across Census Enumeration Blocks (CEBs). Each of the segments were of equal sizes and not exceeding 2000 by grouping these contiguous CEBs. In recent SRS surveys categories of towns/ cities in urban areas are divided into four strata based on population size: Stratum I-towns with population less than 1 lakh; Stratum II- towns/cities with population more than 1 lakh but less than 5 lakhs, Stratum III-towns/cities with population of 5 lakhs or more; Stratum IV- metro cities (Delhi, Mumbai, Chennai and Kolkata). The CEBs within each size stratum were ordered by female literacy and sub-divided into 3 strata. One CEB from each sub-strata was selected using simple random sampling without replacement technique.

LASI

Like most of the large-scale surveys LASI will follow a multi-stage stratified sampling. It will follow the respondents over the survey span to provide cohort data.

NSS

NSS adopts a multistage stratified sampling design. The list of Census villages for rural areas and Urban Frame Survey blocks for the urban areas for the first stage units (FSU) in NSS. In case of large FSUs, hamlet groups in rural areas and sub-blocks in urban areas are the intermediate strata. Sample villages from rural areas and sub-stratum/stratum for urban sector are selected by probability proportional to size with replacement (PPSWR). Simple Random Sampling without replacement (SRSWOR) is used to select the ultimate stage units, households.

NSS draws two types of sample list: central sample and state sample. NSS previously collected data to provide estimates at the national level only. However, with growing needs of sub-regional estimates, state level samples are collected to provide reliable estimates at state-level and for regions within the states.

GYTS

For GYTS, the sample is selected in two stages. At the first stage, schools are selected through systematic random sampling using probability proportional to the school enrollment size. At the second stage, within the selected schools, classes are chosen through systematic random sampling. All students of the selected classes are eligible to participate. The sample frame for GYTS includes both public and private schools and classes of students in the age group 13-15 years. The survey uses Unified-District Information on School Education (U-DISE) data of 2017-18 for sampling.

IHDS

IHDS follows a multi-stage stratified sampling design like other large-scale surveys like NFHS. However, since it is a panel study, the survey re-interviews 85% of the households from preceding rounds. Additionally, they also draw a random sample of PSUs and households using PPS. Comparing the panel sample with this randomly selected refresher sample allows to determine whether the panel sample is overrepresented among certain segments of the society.

IBBS

IBBS uses two types of probability-based cluster sampling methods to sample HRGs from hotspots or “any identifiable location where respondent group members congregate” or “are known to be associated with”.

1. Convention cluster sampling (CCS): To recruit respondents from usual clusters including sites and establishments to which HRG members were affiliated to (eg: homes, brothels).
2. Time location cluster sampling (TLCS): To recruit mobile respondents such as street based FSW, MSM or IDUs from time location clusters. Such hotspots were classified into four categories based on data collected during sampling frame development exercise: peak day-peak time, peak day-lean time, lean day-lean time or lean day-peak time.

6. Sampling units

The main aim of the sampling design is to arrive at the sampling units required to provide reliable and robust estimates of the variables of interest. As a part of the sampling design, large-scale surveys define their sampling units and primary stage units to facilitate effective data collection. Most surveys select the sampling units from clusters of units with low variability within the clusters and maximum variability between the clusters. However, selection of sampling units can be challenging when the survey attempts to provide continuous estimates or when sampling units vary for different sections of the survey.

Examples:

SRS

The SRS sampling frame is revised every 10 years along with the decadal Census of India. During the revision, modifications are made to the sampling design along with increase in population coverage to meet additional requirements. House list, household schedule and list of women in reproductive ages are updated too. Previously, replacement of samples took place in phases over 2-3 years. However, from 2004 onwards, the replacement was done at one go within a year. Since IMR is the decisive indicator for sample size estimation, SRS has set the permissible error level to 15 prse (percentage of relative standard error) at the level of national divisions for major states and at the state level for smaller states.

IBBS

IBBS aims to collect information on high-risk behavior related to HIV among HRG. Hence, the survey adopted a probability-based design to provide representative estimates at different geographical levels. The basic unit was a domain which was usually a district. In case a district did not have adequate number of HRG then neighboring districts were grouped to form a composite domain. Districts were randomly selected for the inclusion in the survey

with specified lower limit of the different categories for HRG, as for example, have at least 800 FSW and MSM and at least 600 IDUs. Districts are further stratified into 3 groups: low, medium and high based on the HRG population size. To ensure representation of all categories, domains within each state were grouped into clusters based on natural divisions (socio-cultural regions as per census, or administrative divisions). One domain from each cluster was then selected.

The sample size for behavioral aspects of HRG is calculated based on:

(i) Expected value of consistent use of safe practices (condoms with commercial partners/ regular partners or clean needles/ syringe) at baseline- 50%.

(ii) Change to be detected at the end line of NACP IV- 15 percentage points

(iii) Design effect- 1.7

(iv) Desired level of significance of estimates (Type 1 error)- 95%

(v) Desire level of power of estimates (Type 2 error)- 90%

(vi) Sample size required- 385~ 400

For reliable HIV prevalence estimates among the HRG, the baseline prevalence and desired change differed for different groups. Thus, sample sizes were also different for the different groups. These were aggregated for individual states or group of states.

7. Selection and training of field investigators

This is a vital step to ensure data quality is maintained throughout the field work. Selecting capable field investigators and training them to understand the questions, techniques of probing, protocols to maintain while asking sensitive questions, proper translations of questions if asked in vernaculars are essential to have proper data. Different surveys have different procedures of selecting field investigators and invest different time period to train these investigators. The technique of training also varies by type of questions and measures that are to be taken on field. The mode of data collection also varies by the survey budget and planning. However, in recent times, majority of the large-scale surveys have started Computer-Assisted Personal Interviews (CAPI) which is a major step to improve data quality as it prevents implausible answers to be recorded through various system algorithms. All training programmes prior to any survey has to be standardized in order to provide a similar footing to all the selected investigators.

Examples:

NFHS

Supervisors are evaluated intensively before appointment based on their education, experience and performance in the test conducted by IIPS, rigorous 3 week long training of trainers of the 4 core team members of field agency (One Social Scientist, One Health Coordinator, One IT coordinator and Demographer)- one ToT for each phase, state level investigator trainings were closely supervised and monitored by PIs, SPO/ PI, ICF and officials from MoHFW, more than 1 training to be conducted for states with more than 10 districts, skill test conducted for investigators.

CNNS

The survey implementors hired 900 interviewers. Around 400 supervisors and quality control staff were engaged during the data collection. One interviewing team consisted of four survey interviewers, two health investigators and two laboratory technicians. The set target was to complete interviews and measurements of 60 children in 5 working days, and to complete the designated number of interviews in a state in 3-4 months. Before data collection all interviewers had a four weeklong training to make them adept to various survey objectives, field practices and procedures for interviews and biological sample collection. Data was collected using Computer Assisted Personal Interviewing (CAPI) method. In order to have reliable investigators, CNNS employed university graduates with previous experience of survey data collection. The supervisors were in charge of overall planning, implementation, coordination and execution of fieldwork, reporting of completion of data collection in each PSU, ensuring that all selected households were visited, and all eligible respondents were contacted. They were responsible to check that the list of household members provided by the interviewers were matching with the original sample listing. The educational qualification to be eligible to be appointed as a supervisor was post-graduation with extensive experience in managing field surveys and agencies.

Since, the major objective of the survey was to provide robust nutrition related estimates, 360 individuals were appointed to collect anthropometric measurements. The requisite for appointment was previous training in public health. During their training, anthropometric standardization exercises were practices to measure their skill set, precision, accuracy. The standardization exercises were performed (except for weight) on 10 children of age 2-10 years. Measurements were recorded twice with a gap of 1 hour between the two recordings. The measures taken by the trainees were compared to a gold standard measurement taken by an expert. Deviant measures were indicated by '+' (for higher than gold standard) and '-'

(for lower than gold standard). These were used to calculate the inter-observer technical error of measurement (TEM) and reliability of the measures. Trainees with unacceptable scores were either dismissed or retrained.

The responsibility of collection and testing of biological samples for the survey was given to SRL Limited who hired 360 experience and well-trained phlebotomists with at least a Diploma in Medical Laboratory Technology for the task.

SRS

The supervisor of SRS belongs to the statistical cadre of the State Census Directorates (either a Compiler or a Senior Compiler or a Statistical Investigator or any suitable official). The SRS also publishes special bulletin on maternal mortality and causes of death statistics. Causes of death is collected using post-Death verbal autopsy method by conducting post death enquiries on the family members' observations on symptoms, conditions, duration and anatomical site of the disease of the deceased. Supervisors of SRS are trained by medical professionals to comprehend the medical terminologies, symptoms and conditions related to causes of death. To maintain the quality and objectivity of the data, supervisors are recommended to strictly follow the verbal autopsy instrument and interview close relatives of the deceased. The supervisors are trained to fill the objective closed questions and the narrative portion of the verbal autopsy instrument.

LASI

ToT workshops were carried out in two stages and survey schedules were translated into 18 languages. The PRC researchers were trained at the first meeting and the trained the interview teams at the state level. Like other surveys, in order to strengthen interviewer-interviewee rapport and provide a conducive environment to ask sensitive questions and

conduct biomarker measurements, gender matching was done to ensure that the interviewer was of the same gender as the respondent.

IHDS

Extensive training was provided to field teams by the NCAER staff and researchers from University of Maryland. Eleven 2-week log training programmes were arranged across the country for 15-50 interviewers. Classroom reviews of questionnaire along with supervised field experience was also given to the field investigators. Along with the written interviewer manuals, short films on data collection procedure and probing were developed for easy understanding of the interviewers.

IBBS

During the training period, all laboratories, field agencies and investigators were made aware of the standard guidelines and training materials that were developed to address different needs and sub-categories of HRGs. IBBS prepared a detailed Interviewer Manuals and Field Laboratory Manuals to guarantee collection of good quality data. Trainings were standardized and training agenda was systematically developed to strengthen capacity and capabilities of the survey teams. The field teams were trained through a cascade of training programmes:

1. Training for pre-testing
2. National level training of trainers (ToT)
3. National level training on IT component
4. Regional level training of investigators for 1 week by Regional Institutes in charge
5. Two-week field level training in each state

8. Pre-test

The final questionnaire is developed only after proper pre-test before the full-fledged field implementation. Pre-test or a pilot survey is carried out by administering the entire questionnaire to a small group of target respondents in a designated setting as decided in the initial stages of planning. It is an essential tool to maintain data quality standards. It helps in revising questions that are unclear to respondents or may generate erroneous responses. Pre-test also helps to understand the average time being taken to administer the entire survey questionnaire and thereby have achievable timelines.

Examples:

CNNS

Before pretesting, the questions were translated into 20 state-specific languages to maintain easy comprehension and response to the questions. The questions were pretested on 100 respondents in the rural, urban and slum areas of Delhi, Maharastra, Uttar Pradesh and West Bengal. The questions were revised based on feedback on social and cultural acceptability, complexity of questions, rational order and structure of the questionnaire.

Further, a comprehensive pilot test was conducted on 240 respondents in four PSUs in Bihar in the two weeks prior to the survey. The entire survey was simulated for the selected participants and final calls on data collection and survey implementation was carried out based on the findings from this pilot survey.

LASI

LASI conducted a pilot survey on 1600 individuals administering the entire questionnaire. The pilot data was published and has been used in several research works. The pilot data was collected through face-to-face interviews over three months in 4 states: Punjab, Rajasthan,

Karnataka and Kerala to comprehend regional variation that is predominant in most indicators of public health in India.

GYTS

In case of GYTS, any designated location was selected to conduct pre-test on six to eight 13-15 year old participants. After debriefing and administering the questionnaire using the standard protocols, the facilitator collected feedback from the participants. This will help in revising questions that were ambiguous or unclear to the respondents and yield more accurate data collection during the field implementation.

IBBS

Since IBBS collected an array of biological and behavior data for the HRGs, more than 20 different tools and formats were developed. The questionnaires were translated to several local languages. Tools used in sampling frame development and informed consent form were pretested using standard protocols using a hard copy of questionnaire as well as CAPI.

Chapter 3. Data quality checks during the survey

This chapter will document some of the data quality assurance measures that are taken by various large-scale surveys during the fieldwork. Most large-scale survey employ a field feedback process through field supervisors, regular monitoring of field data and evaluation of data that is being collected. Regular monitoring and expert supervision are required to ensure quality of various biometric measures. Often some experience project staff is appointed as a field supervisor to ensure that investigators are interviewing in the recommended manner to guarantee minimum non-sampling error arising from the data that is collected. In the following examples from different surveys, we list out some of the data quality protocols followed during a survey pertaining to fieldwork organization, on-field monitoring and evaluation and feedback from supervisors.

NFHS

1. Data synchronization between investigator and supervisor's CAPI provides opportunity for retracing errors and revisiting discrepancies
2. CAPI allows for partial saving of interviews thereby facilitating multiple visits in case of respondent fatigue or unwillingness to answer the entire questionnaire in one seating
3. SyncCloud Technology synchronizes data between supervisor's CAPI and Central Office, which helps in accessing data in real time on a regular basis
4. Field check tables help in recording and providing regular feedback
5. Supervisors can generate error messages before completion of survey in a PSU if there are any internal inconsistency observed. In previous rounds, FA was responsible for secondary editing of data before final submission. These new changes reduce the task of secondary editing by FA's upon completion of data collection. If inconsistency is

observed, then the team supervisor may ask the investigators to revisit some respondents and thus, FA's effects on data editing is nullified.

6. Supervisors can generate error messages before completion of survey in a PSU if there are any internal inconsistency observed. In previous rounds, FA was responsible for secondary editing of data before final submission. These new changes reduce the task of secondary editing by FA's upon completion of data collection. If inconsistency is observed, then the team supervisor may ask the investigators to revisit some respondents and thus, FA's effects on data editing is nullified.
7. Project Officer's Query on Supervisor's CAPI (POQR) has been developed. Upon completion of data collection, this algorithm is effective in retrospective checks by the Project Officer at IIPS who can run the tool to identify households with data issues in a PSU. This has made back checking and revisiting households with data inconsistencies smoother in the recent round of NFHS.

CNNS

India was divided into four zones, with seven to eight states in each zone. Five to six field teams were employed to work in each state. In each of the zones, data was collected in two states simultaneously to ensure balance workload among the four survey agencies.

In order to maintain data quality during the field work, performance of interviewers was supervised at three tiers.

In the first tier, quality control staff monitored and observed the interviews, respondent interviewer interactions and provided feedback to the investigators for improvement of interviews. Almost 30% of the interviews in each PSU was monitored and regular feedback was provided to improve interview techniques. Each zone and state had a zonal field manager and state field manager who were responsible for training, supervising, monitoring and

providing feedback to the data collectors. They provided feedback in revising interview techniques based on their observations of 2-3 interviews in each PSU in the state and zone.

In the second level, a three member data quality assurance team monitored and supervised 80% of the PSUs in a state and visited 10% of the households in each PSU to validate the household listing process. If the inconsistencies were higher than 10%, listing was redone. This team observed three interviews in the PSUs they visited and re-conducted a minimum of three interviews for quality assurance. The survey used the computer-assisted field editing (CAFE) tool of CSPro to assess the data for the re-interviewed questions and provide appropriate feedback as a data quality assurance protocol.

The data quality checks in the third tier were done by experts from PGIMER, UNICEF and Population Council by observing interviews and providing feedback and corrective measures at different stages of implementation. Every 15 days, field check tables were generated and reverted back to field supervisors for reference.

In case of anthropometric measures, two trained anthropometrists worked as a team to record all the measurements. Other components to warrant good quality of the anthropometric measures such as lighting present in the place of measurement, availability of flat surfaces, etc were also taken into account during data collection. The similar three tier data quality assurance protocol was adopted for anthropometric measures too. The quality control staff and supervisors monitored the calibration of the equipment, observed a proportion of the measurements being recorded in the field and provided feedback or sent back the anthropometrists to re-take the measures if found unsatisfactory.

Biological samples were collected the morning after the anthropometric measurements to ensure fasting samples. The SRL Limited was entrusted with safe collection, packing and transportation of these samples. Standard operating procedures were put in place to maintain

high quality of the biomarker estimates. All the equipment were validated on a regular basis. For this case too, three levels of quality assurance protocols were in place. First, each batch of 20 survey samples were subjected to internal quality control. Second, a subset of the samples were re-examined by a separate participating laboratory to validate the results every month. Third, samples were split and re-tested on a weekly basis. All India Institute for Medical Sciences (AIIMS) laboratories were also engaged in validating by re-assessing subsamples for microscopy. Further, 5% of the samples were randomly selected and tested at the National Institute of Nutrition (NIN) and AIIMS laboratories. Regular visits were made to the laboratories by the United States Centre for Disease Control (US CDC) and the Clinical Development Services Agency (CDSA) team was responsible for supervision of collection of the biological samples.

SRS

The SRS in India is based on the dual record system. In the first level, the on-field continuous enumeration of births and deaths in a sample village/ urban block is carried out by a resident/ local part-time enumerator. In the second level, an independent half yearly retrospective survey (January-June or July-December) is conducted by a full-time supervisor. The supervisor does not have access to the births and deaths enumerated by the investigators before the supervisor's field visit. The data discrepancies that emerge from matching these two sources are used to address data quality issues in the survey. The staff (either a third person or jointly by the supervisor and enumerator, depending on availability) re-verifies the unmatched and partially matched events in the field to get an unduplicated and accurate count of the true events. The items and events that are used for matching are: Identification code of the head of Household and mother, Relationship of the mother to head, date of live birth, month in case of still birth/abortion, sex in case of live birth /still birth and the items considered for death events are: identification code of the head of household and mother in

infant death, relationship of the deceased to head, date of death and sex of the deceased. The benefit of this process is not only elimination of errors and duplications but also promotes a self-evaluating technique by giving scope to the implementors to perform a quantitative assessment of the two sources.

LASI

Multi-stage quality control was administered. Like CNNS, LASI too had a three-tier supervision and monitoring by field supervisor, research officer at the state level and project officer in the nodal agency.

IHDS

For IHDS I, 25 agencies were employed based on their experience in handing large-scale surveys. The field teams consisted of 5 members: 2 pairs of male and females' interviewers and a team leader. The team leader was responsible for conducting the village, school and medical facility interviews and supervise the interviewers. The field investigators were supervised through random visits by supervisors and zonal coordinators from NCAER and 2.5% households in each district were partially re-interviewed by NCAER's own field staff in each state. Editing staff at NCAER were responsible for identifying issues of missing data, skip patterns and coding. Such multilevel reviews helped in strengthening data quality and enabled prompt identification of problems and speedy feedback process. IHDS II augmented this field monitoring technique by introducing time/place stamp using a GPS enabled phone. Interviewers were directed to upload field data in the central server daily to facilitate smooth monitoring and supervision during field work.

IBBS

To ensure intensive project management and smooth implementation of the survey a web-enabled Integrated Information Management System (IIMS) was installed. The IIMS was accessible to different management level staff and provided timely information on key aspects of human resource management, supply chain management, progress, field and laboratory database management, monitoring, grievance redressal and adverse event management. All the data collected via CAPI was automatically synced with IIMS.

Like other large-scale surveys, intensive monitoring and supervision was also a part of IBBS. Both internal and external monitoring and supervision was in place at different levels. The Regional Nodal Institutes and NWG were responsible for external monitoring to ensure strict adherence to standardized guidelines and protocols. All field teams were visited within 15 days of commencement of the main field study. IIMS was used for effective real-time monitoring. A comprehensive supervision and monitoring framework describing roles and responsibilities of all engaged agencies was specified.

National AIDS Research Institute (NARI), Pune was the designated Apex Laboratory for IBBS. The ensured proper biological testing in the assigned regional laboratories through quality assurance protocols and External Quality Assessment (EQA). Proficiency assessment of the laboratories through panel testing and re-testing was conducted. Retesting of all positive samples and 2% of the negative samples was conducted at NARI. Standard Operating Procedures (SOPs), rigorous training for DBS collection, storage and transportation was also carried out in details.

Chapter 4. Post-survey data quality checks

The fourth chapter will aim to discuss what are the post-survey measures that are employed to ensure the data that is published for large-scale public use is reliable and usable. After the entire fieldwork process, the raw data is aligned in an intelligible manner for ease of use by public. The target audience of surveys are usually researchers, policy makers, scholars, bureaucrats, funding agencies. To help easy use of the data, the data needs to be cleaned and all data with implausible inputs are to be recoded in appropriate manner. It is the responsibility of the survey coordinators to provide all data maintaining the ethical considerations of the survey. At the end of the survey, a detailed report, explaining how each indicator has been generated along with detailed definitions is published to provide an overview of the indicators available in the survey. Dissemination of data along with all the data quality pointers is an essential part of any large-scale survey.

(i) Cleaning of data

Data cleaning is a major part of ensuring quality data provided by the survey for public use. For this purpose, the data entry operators have to take special care to flag implausible cases and mark missing responses. Post data collection, based on the sampling design, many surveys generate sampling errors and design effects for various important estimates. This is usually used by administrative bodies to check the confidence that can be bestowed upon the estimates. For example, NFHS, CNNS provides a detailed standard error, confidence interval and design effect for most of the important indicators as a part of the appendix in their reports.

(ii) Generating survey report

It is important to maintain a pattern while publishing the report and also highlighting the essential indicators from the survey by various background characteristics of the respondents. A good report must entail a brief but useful introduction to the objectives of the survey along with the need of having a particular survey. The methodology for calculating all the indicators from raw data must be presented in the report to facilitate replication and easy use of the raw data. Standardized estimations of popular indicators help in comparing surveys. Disbursing findings from the survey should be a public event so that users of the data are aware of the various aspects the survey covers. Most surveys do not discuss about various data protocols followed during the survey. It might be a good idea to come up with a supplementary report on data quality protocols and challenges during the survey to strengthen trust and reliability among the users. In the following examples, we describe some of the post survey practices followed by a few large-scale surveys in India.

Examples:

NFHS

1. Data for assessing quality of survey is often missing post-survey. Surveys should make tools for data quality analysis available for public use.
2. One of the good practices followed by the DHS team is regular research on various data quality aspects of already published data.
3. Referring to the analyses used in these reports, future large-scale surveys like NFHS can collect more data on interviewer's characteristics, experience, sex, whether he/she was local or outsider, use of translator in the interviews, time and day of data collection. Such variables, if made available in public domain, can be used for in

depth data quality analysis of these surveys. At present, only a limited number of research papers deal with data quality of indicators in NFHS and most of these works are argumentative rather than being mathematical.

IHDS

Data entry and cleaning was centralized at NCAER, New Delhi. The questionnaire was mostly self-coded for the ease of data entry. The report estimates were checked for consistency during data entry and report writing and troubleshooting was done by rechecking original questionnaires and via phone calls to interview site/ field agencies.

Chapter 5. Best practices adopted in surveys and learnings for future

The last chapter will provide a detailed assessment of how various large-scale surveys can improve data quality and provide some expert note on the way forward. Due to the present pandemic situation, field interviews might be a challenge. This chapter will provide some insights on what are the data quality considerations that are the need of the hour and how can surveys transition in view of the "new-normal".

Government organizations and departments in India often collaborate with international and national agencies to collect data on multiple dimensions for policy reformations and program planning. To bring awareness about actual demographic, socio-economic and health situation, high quality data is crucial for responsive health systems. These data often suffer from quality threats for various reasons such as sampling and non-sampling errors. With the growing investment in collection of various kinds of data in the country different strategies has been adopted to maintain data quality.

One of the good practices followed by the DHS team is regular research on various data quality aspects of already published data. This helps identify areas where quality improvement is required. These research on data quality is published as methodical reports by DHS on their official website and usually talks about age heaping, age displacement, data quality of fertility indicators, anthropometric measures, interviewer bias in data collection, etc. Most of these are cross country research work which involves several DHS countries subjected to data availability. Referring to the analyses used in these reports, future NFHS can collect more data on interviewer's characteristics, experience, sex, whether he/she was local or outsider, use of translator in the interviews, time and day of data collection. Such variables, if made available in public domain, can be used for in depth data quality analysis

of these surveys. At present, only a limited number of research papers deal with data quality of indicators in NFHS.

It is ideal for large-scale surveys to self-evaluate using preceding rounds to upgrade their research methodology and sampling techniques and frame. For example, SRS revises the sampling frame every 10 years and modifies their sample design and techniques of data collection based on previous surveys. Like before 2004, the survey used birth rate to decide the sample size but post 2004, the survey is using IMR to decide the sample size.

Innovation in survey instruments is the need of the hour to ensure top quality data. For example, the DIID of NSO is planning on building a state-of-the-art digital repository called National Integrated Information Platform (NIIP) for assimilating the official statistics with homogenized meta-data for India. The designs of surveys need to be reconsidered and transformed in the light of the ongoing pandemic situation where social distancing and limited human interaction has become the regular norm. Redesigning surveys to ensure that data collection is continuous and within expected timeline has become very challenging in recent times. Taking this into account a few are thinking of innovative ways of data collection. For example, DQAD of NSO is planning to develop a Generalised Survey Solution namely e-SIGMA, (e-Survey Instrument and Generalised Multimodal Application) for the surveys conducted by NSO. This application will encapsulate all the facets of the survey in one place and will become the one stop solution for all the divisions under NSO. This will allow for seamless transition of data from the field to the DQAD for further processing. This software will have the facility to capture data through the CAPI module and provide Real Time Validation through the Back Office Module. This will also decrease the time of processing and publishing data. Along with these novel efforts, call-centre based survey options is also being set up in DQAD in the light of the new normal due to COVID 19 pandemic. These call centres will be equipped to conduct remote surveys through CATI

(Computer Aided Telephonic Interviewing) and CAWI (Computer Aided Web Interviewing) using IVRS (Interactive Voice Response Services) technologies.

Some of the best practices that are adopted in a few surveys to ensure better quality data in large-scale surveys are listed below:

- All the standard data surveys secure the educational qualifications of each field and office staff and their educational qualifications are verified by central office before start of the work.
- Mapping and listing is one of the important step to prepare the sampling frame for household selection. To deal with real on-field issues, field practice would be conducted after providing quality training with limited participants which will reduce the chances of non-sampling error. Survey coordinators and senior staff checked the data quality during field practice and if found inadequate additional training was provided.
- In most of the surveys data needs to be collected from both male and female respondents. It has been observed that female respondents feel more comfortable opening up to female investigators than males. Therefore, to ensure high quality data, surveys keep a gender sensitive team arrangement with a blend of an adequate number of male and female investigators who have expertise in the local language(s).
- Almost all the surveys have shifted from pen and *paper interviewing* (*PAPI*) *to computer-assisted personal interviewing* (*CAPI*) which helps in eliminating errors by automatically following skip patterns and filters.
- In large scale databases, to maintain data quality training were provided at all levels (state and central) to the survey team members and central team were responsible in overseeing the data collection process. Supervisors provided their feedback on a time to time basis and were responsible for reporting survey progress to their respective offices.
- To strengthen the data quality, survey central office randomly selected 10% Primary Sampling Unit (PSU) for spot check and back check surveyed households in addition to regular review of field check tables.
- According to the survey (NFHS) protocol, 10% of the surveyed households were back checked by supervisors in which they cross check the height and weight of the children to maintain quality of data.

- In central office, real time field check table has been generated to check for birth displacement and to ensure skip pattern has been adopted properly.
- After mapping, listing and segmentation, selection of segments as well as households has been done at central office by survey coordinators.
- Project officers (PO) along with supervisors were involved in allotment of selected households to each interviewer and evaluated the data quality under the supervision of the project co-ordinators.
- All the interviews are checked by the supervisors/ Project officer before sending them to the central office.
- In some surveys like LASI, as per the data quality protocol, there was a presence of central office (IIPS) in the form of a PO for back checking in every selected PSU.
- On a timely basis random questions were generated from any section of the questionnaire for back check.
- For ease of management, multiple PSU was selected within selected tehsil.

After reviewing some of the most popular surveys, we have listed out some of challenges that need to be addressed to improve data quality in existing large-scale surveys in India.

(i) Respondent based bias remains an issue of concern in most demographic, health and nutrition surveys. Surveys should develop frameworks to ensure respondent engagement and commitment to the survey.

(ii) Varying efficiency of field agencies and investigators often affect data quality. Before the start of field work investigating teams should be evaluated based on collection of pilot data to understand the effectiveness of the training imparted.

(iii) Unstandardized reference periods of different estimates make it difficult to compare similar indicators across surveys.

(iv) Most samples collect an array of variables among which a few remain unutilized due to poor quality data. This can be handled by reviewing the usage of variables in research and dropping out variables or modifying instruments of collection based on the usage of data.

(v) Surveys collecting similar information often adopts different scales of measurement. A lack of standard length of questionnaires often makes if difficult to compare survey instruments' validity across surveys.

(vi) Lack of standardization of definitions also makes indicators incomparable.

(vii) Academic involvement in the form of course curriculum on survey design and research can not only increase awareness of various tools and techniques used in survey data collection but also promote innovative survey approaches.

(viii) As we are entering into the big data era technical innovations in surveys are the need of the hour.

(ix) In-depth research should be promoted to develop methodologies and indirect tools to provide accurate estimates of various sensitive statistics that often suffer from poor data quality in demographic, health and nutrition surveys.

(x) Customization of questionnaires and research instruments to collect accurate data from different sub-sections of the sample should be critically supervised and evaluated.

(xi) Follow-up surveys/ ad-hoc surveys can aid in better data validation and revision of survey methods.

(xii) In the light of the recent pandemic, surveys should come up with a plan B to ensure that data collection is not disrupted, and quality of data remains high. This can be done by investing in remote survey designs.

Figure 3. Instruments to produce good data in large-scale surveys