# GUIDELINES ON STANDARDISING DATA COLLECTION

## 1. INTRODUCTION

Data accuracy is essential to maintaining the integrity of research and for reporting of results. To reduce the likelihood of errors,[1] it is critical to select the suitable instruments for data collection and ensure that they contain clear instructions for their correct use. The extent of the impact from faulty data collection varies, with the biggest risk being causing disproportionate harm if such data is used to support policy recommendations.

TMEA will standardise all its data collection activities to reduce errors, ensure consistency and comparability of the reported results and avert any other potential unforeseen risks. The standardisation is critical in tracking and measuring performance of TMEA's indicators which are benchmarked on its core outcomes and outputs in the approved Strategy II Results Framework.

The Research and Knowledge (RK) Unit has developed guidelines for teams to observe with the goal of ***quality assuring and controlling*** the entire data collection process. This will ensure that the data collection processes have scientific rigour, provide data consistency and can be compared across programmes and projects. These measures will preserve data integrity (by reducing chances of errors i.e., deliberate falsifications or systematic or random errors) and thus guaranteeing reliability and credibility in analysis and reporting of the results. Additionally, they are important recipes for data transparency and replications (if needed).

## 2. OBJECTIVES OF THE DATA COLLECTION PLAN

### *2.1 Purpose*

1) To provide evidence on the specific impact of TMEA's interventions across projects and programmes.
2) Contribute to the collection of micro-data that adds to the broader debate on the trade - development nexus that has dominated policy discourse over the last six decades.
3) To make TMEA a data hub on trade facilitation in Africa region and beyond.

Towards these objectives, TMEA aims to collect data that respond to its research and reporting needs:

**a) Data collection for ex-ante analysis**

This entails gathering information that directly feed into the IMPACT model. The primary goal is to; (i) calibrate model parameters to reflect the reality on the ground (ii) provide the basis for decision making by forecasting the investments of highest returns with the possibility of the highest welfare impacts.

[1] The costs of improperly collected data include: failure to answer research/policy questions accurately, failure to replicate and validate the study, distorted findings ensuing in misused resources, misleading other researchers and practitioners to pursue futile avenues of research and policy investigation, compromising decisions for evidence-based policy making, and potentially causing harm to human subjects.

# GUIDELINES ON STANDARDISING DATA COLLECTION

**b) Data collection for ex-post analysis**

This relates to the collection of data that meets the objectives of specific project's needs. It entails both baseline and end-line survey data collected before and after the implementation of the various interventions.

In both ex-ante and ex-post analysis, the goal is to collect primary data (mostly baseline data) and where necessary, secondary data that complements our focus to measure some of the indicators.

## 3. BEST PRACTICE FOR DATA COLLECTION

As part of the best practices, TMEA's data collection activities constitute two key components, *quality assurance* and *quality control*, as described below:

### 3.1 QUALITY ASSURANCE (QA)

Quality assurance includes activities that take place ***before*** data collection begins. This entails anticipating potential problems with planned data collection activities. The key element for quality assurance is problem prevention which is the most cost-effective way to ensure data integrity and guarantee reliable and high-quality data. This proactive measure involves the standardisation of data collection protocols.[2]

The following selection criteria help in defining the nature and quality of the data collection:

**a) Reliability:** This entails a detailed thinking on the consistency of measurement and whether the proposed measures/variables represent the core concepts that the data collection aims to achieve. Most importantly, it involves ensuring that the methods used for data collection are accurate and transparent.

**b) Validity:** This is divided into three parts; ***internal validity*** which focuses on whether the relationships between variables of interest is causal or not; ***external validity*** which weighs whether the results from data collection can be generalised beyond their specific context; and ***ecological validity*** which determines the relevance, practicality and applicability of research findings.

**c) Competence:** Developing a rigorous and detailed recruitment and training plan of field teams to effectively communicate to them the value of accurate data collection.

### 3.2 QUALITY CONTROL (QC)

Quality control includes all activities (detection/monitoring, verification and action) that take place ***during*** and ***after*** data collection. To monitor data, teams and consultants must make sure that all practical details and procedures related to data collection are precisely documented in the data manual/protocol and plainly communicated to data collection teams.

The manual/protocol must also lay down clear reporting lines (between principal investigators and enumerators) during the field work. TMEA's project staff (jointly with the data team of the RK Unit) must

[2] These include detailed and comprehensive manuals and procedures for data collection.

ensure that data collection protocol/manual is strictly adhered to. This can take the form of observations during field visits, or regular and frequent reviews of incoming data reports to flag any data inconsistencies.

Follow up queries to respondents, about the collected information, can be done in various ways and will further control data quality. For example, teams/consultant can do cross-checks within the data collection process and make observations to support what's being said. Data quality should be addressed for each individual measurement, observation, and for the entire data set. Finally, TMEA staff and/or the consultant must jointly put in place the actions necessary to correct faulty data collection practices[3] and minimise future occurrences.

Thus, as part of both quality assurance and control, all teams must evaluate the type of data needed for analysis and reporting. Most often, teams can collect two types of data: primary and secondary. The focus of this note, however, is on primary data – the baseline data collection – which is the central focus for evaluating outputs and outcomes of TMEA's Strategy II.

The next subsections provide a brief description of methodological awareness (i.e., research/study design, data collection, and data analysis) and a recap of practical data management considerations before (**i.e., quality assurance**), during and after (**i.e. quality control**) data collection activities.

### 3.3 BEFORE DATA COLLECTION

Before commencing on any data collection, ensure that the RK Unit has reviewed and cleared the data collection proposals. Additionally, teams should pay attention to the following:

**a) Define the general area of interest and develop the specific questions/hypotheses that guide data collection:**

i) How project activities speak to TMEA's broader strategic and corporate goals.
ii) The description of the questions/hypotheses should be clear, comprehensible, and precise.
iii) Consider the potential links of the proposed interventions to gender and climate change (if any).

**b) Determine the design of the study and data collection by paying attention to:**

i) **Scale:** units of analysis (individual, team, firm or industry) and the number of units. The decision on the sample size should be scientifically made (using credible and reliable statistical methods) to address issues of representativeness and avoid potential biases. Note that sample size is an essential ingredient for developing data collection budget(s).

[3] Examples of data collection problems that require prompt action include: errors in individual data items, systematic errors, violation of protocol, problems with individual staff or field performance, fraud or scientific misconduct.

# GUIDELINES ON STANDARDISING DATA COLLECTION

Teams or the consultants must make sure that a detailed and scientific sampling procedure is included in the ToR for data collection. **The recommended approach is for teams to use the probabilistic sampling techniques**. Depending on the nature of the data collection exercise probabilistic sampling techniques can include the ***simple random sampling (SRS)[4], systematic[5] sampling, stratified[6] sampling, and cluster[7] sampling***.

ii) **Time:** Will the study focus on a single point in time or is repeated over several periods of time (for which it is useful to choose the measurement frequency).

iii) **Comparative dimension:** Will the study compare outcomes over time and across different units of interest.

iv) **Type of data:** Experimental data (collected on a specific experiment), cross-sectional data (collected over many units of study but only at a single point in time), longitudinal data (collected over many units over time), and case study data.

v) **Data instruments:** Designing appropriate data collection instruments (e.g. questionnaires that address the core questions/hypotheses underlying the data collection exercise and address issues related to gender, environment, and climate change whenever necessary – See **Annexe 1**)

vi) **Protection of human research/data subjects:[8]** Teams (and/or the consultants) must always adhere to the standard protocols and ethical practices that guide the protection of human subjects throughout data collection process. The hired consultants **must be certified by relevant national and international bodies[9]** on collecting data on human subjects. This goes beyond strictly observing all the rules, regulations and ethical practices related to surveying and interviewing human subjects. Project teams should make this part of the evaluation criteria.

All these factors have a significant bearing on the budgets that teams develop or commission consultants to prepare for data collection activities.

[4] This is sampling technique in which each member of the population has an equal chance of being selected in the data collection sample. Simple random sampling allows the sampling error to be calculated and reduces selection bias.

[5] Individuals are selected at regular intervals from the sampling frame. The intervals are chosen to ensure an adequate sample size. Systematic sampling is often more convenient than simple random sampling, and it is easy to administer. However, it may also lead to bias, for example if there are underlying patterns in the order of the individuals in the sampling frame, such that the sampling technique coincides with the periodicity of the underlying pattern.

[6] In this method, the population is first divided into subgroups (or strata) who all share a similar characteristic. It is used when we might reasonably expect the measurement of interest to vary between the different subgroups, and we want to ensure representation from all the subgroups. Stratified sampling improves the accuracy and representativeness of the results by reducing sampling bias. However, it requires knowledge of the appropriate characteristics of the sampling frame (the details of which are not always available), and it can be difficult to decide which characteristic(s) to stratify by.

[7] In a clustered sample, subgroups of the population are used as the sampling unit, rather than individuals. The population is divided into subgroups, known as clusters, which are randomly selected to be included in the study. Clusters are usually already defined, for example individual groups or towns could be identified as clusters. In single-stage cluster sampling, all members of the chosen clusters are then included in the study. In two-stage cluster sampling, a selection of individuals from each cluster is then randomly selected for inclusion. Clustering should be considered in the analysis. Cluster sampling can be more efficient that simple random sampling, especially where a study takes place over a wide geographical region.

[8] Teams should seek the guidance of SIT team as this could be a part of TMEA's safeguard measures.

[9] An example of an international body includes the National Institute of Health which provides training and certification on human subjects' protections. Others include Global health training network and fhi360.

# GUIDELINES ON STANDARDISING DATA COLLECTION

**c) Deciding on the approach for data collection and analysis:**

i) Quantitative: this **variable-centred** approach entails developing quantitative measures/variables that are specific, measurable, attainable, relevant and timely (i.e. SMART) and directly speak to TMEA's core objectives across projects and at the corporate level.

The selection of quantitative data goes hand in hand with the collection of the appropriate scientific methods for analysing the data.

ii) Qualitative: this **meaning-centred** approach entails collection of informative qualitative data that is rich (with social context side of data) to provide further insights that can explain the numbers coming from the quantitative approaches.

Qualitative data is useful at understanding the "why" of potential impact and can include consistent but unstructured interviews, focus group discussions, and consistent but unstructured observations.

While both types of data collection approaches are useful, **teams should focus more on collecting quantitative data** which is the basis of tracking, measuring and reporting the progress of our outputs and outcomes.

### 3.4 DURING DATA COLLECTION

Collection of data can be in four ways: interviews, observations, questionnaires, and literature search. All teams should always aim at collecting quantitative data that are useful for quantifying impact and it is proposed to ensure that standardisation is implemented across the board, TMEA's baseline data collection should mainly include questionnaire instruments.

Questionnaires are useful and pragmatic tools for collecting reliable and valid data, especially from large samples. It should be mandatory for all teams to submit all their data collection questionnaire instruments for review, approval and clearance with the research and impact team. Adhering to this will guarantee that standardisation is mainstreamed in data collection, by offering higher validity and allowing less flexibility on survey instruments.
Questionnaires can be binary responses (e.g. yes/no), or on a Likert-scale or on a semantic differential scale (e.g. good – bad, sweet – bitter etc.).

# GUIDELINES ON STANDARDISING DATA COLLECTION

To develop concise questionnaires teams should observe, at a minimum, that;

- All questions are understandable
- Phrases are clear, simple and straightforward (affirmative, no double negation)
- Questions are not on multiple dimensions
- Questions are not leading (suggestive)
- They categorise questions into common topics
- Measure constructs with several items to ensure higher reliability
- Whenever possible and available they use existing scales.

### 3.4.1 CAPTURING AND STORING DATA

To ensure consistency in data capturing, all teams are required to use computer-assisted personal interviews (CAPI) – where data is captured using smartphones or tablets – to code all questionnaires. The use of CAPI minimises errors and ensures consistency and timeliness in data collection. For practical purposes, teams should consider using either of the following CAPI tools: (i) the **Census and Survey Processing System (CSPro)** [10] or (ii) the **World Bank's Survey Solution software (SS)**.[11] CSPro is free while the SS is available for free in developing countries.

For secure data storage, teams should consider two alternatives. First, if data collection is done internally, teams should make use of TMEA's servers as a repository of all incoming raw data. As part of data quality control, teams should work with the IT unit to monitor data collection in real time and flag any issues that are likely to happen during the field work.

Second, if data collection is outsourced the consultant must use TMEA's servers as the main repository for all incoming raw data. The consultant should also have their own servers or buy cloud server access (in case they do not have their own servers) which shall be used as mirror servers for backup when data collection is ongoing. Teams must ensure that hired consultants have the experience and technical capacity to collect data using either of the suggested CAPI tools.

### 3.5 AFTER DATA COLLECTION

Teams must confirm that the collected data meets the minimum expected standards before the analysis begins.

[10] Available at https://www.census.gov/data/software/cspro.html

[11] Available at http://surveys.worldbank.org/capi

### *3.5.1 Data Analysis*

Whenever possible teams should get involved in data analysis to ensure correct interpretation/conclusions regarding questions/hypotheses under investigation for reporting purposes. The RK Unit recommends using **STATA or R software** for data analysis. STATA and R are widely applied software for performing statistical analyses in hard and social sciences. The key advantage of both software is that they allow script coding which makes replication of the estimated results a press-of-the-button matter.

Importantly, for ease of replication and checks of the estimated results, teams and/or consultant must ensure the analysis is coded in a coherent and logical manner with clear descriptions of what the codes are doing. The codes must clearly and sequentially show each of the estimated and generated results in the report.

Finally, teams must review all the final reports and where possible present such reports to internal and external stakeholders for comments and feedback.

### *3.5.2 Data storage*

Once data analysis is done and reports are submitted, teams and/or the consultant must ensure that raw data files, cleaned data files and their cleaning codes, data code-books and documentation manuals, and final analytical reports are all securely stored in TMEA's central repository servers. Storing such data must be done categorically and to ensure easy retrieval. All submitted final reports must have a cover page that contains, amongst others, **title, (any subtitles), key words etc.,** that can be used to identify the entire report (including the data collection activities and stored data). **Annex 2** provides a sample of such a cover page.

### *3.5.3 Payment to consultants*

Before final payment to consultants is done, the teams must get a final clearance on the quality of the deliverables from the RK Unit. Teams in conjunction with the RK Unit shall review such codes and verify all the estimated results reported by the consultant before making final payments. The RK Unit reserves the right to recommend non-payment in case the submitted codes do not produce the reported results or if it is discovered that the overall analysis is based on unacceptable and unethical data practices that provide biased results/estimates or if the data has been tweaked to influence the results in some substantive ways that raises alarming questions for analysis and reporting.

## 4. CONCLUSION

This note provides a succinct summary of the key issues that all project teams at the regional and country level must pay attention to before, during and after data collection. Though not exhaustive of all issues that are likely to emanate from data collection activities it provides a basic framework for project teams to adhere to as they embark on data collection endeavours. The note shall be updated regularly, to mirror emerging lessons that are worth documenting.

For further clarification on issues contained here-in, please consult the RK Unit.

# GUIDELINES ON STANDARDISING DATA COLLECTION

**ANNEX 1: SAMPLE TEMPLATE FOR BASIC VARIABLES FOR INCLUSION DURING DATA COLLECTION**

This template contains the minimum information that all teams must seek to capture during data collection. The variables mentioned here are not exhaustive and can somewhat be context specific. Project teams (together with consultants) should endeavour to expand on this list and collect as much information as possible.

- Age of the respondents
- Year and month of birth
- Sex/gender of the respondents (or head of the household if it is a household survey)
- Marital status
- Years of education
- Household size (if married)
- Number of children (alive)
- Educational attainment (primary, secondary, tertiary)
- Literacy (yes/no)
- Disability status (yes/no)
- School attendance (yes/no)
- Employment status (formal employment, self-employment etc.)
- Types of occupation
- Average number of hours worked in a day
- Income (wage, pension, salary, remittances)
- Citizenships and nativity
- Country, region, district of residence
- Rural/urban residence
- Migration status (yes/no)
- Consumption expenditure (food and non-food expenditure)
- Asset ownership (e.g. dwelling, animals, farm, car, cellphones etc.)
- Data on potential environmental and social risks
- GPS location of the place of interview/residence/work (this should be captured by the tablets or smart phones)

# GUIDELINES ON STANDARDISING DATA COLLECTION

**ANNEX 2: SAMPLE COVER PAGE**

To ensure consistency archiving and retrieving information, the collected data and its associated documentations (including preliminary and final reports) must be attached with the following cover page:

- **Project title:** The title of the project
- **Project sub-title:** The sub-title of the project
- **Team leader(s):** Names of TMEA staff involved in overseeing the project
- **Dept/Units:** The name of the Unit(s) that is overseeing the project
- **Country:** Name of the country where the project is being executed. If it is regional project, then should be indicated as such.
- **Budget:** Budgeted amounted for the work to be conducted.
- **Consultant(s):** Names of the consultant (including the firm's name and principal team members)
- **Consultants' country:** Name of the country or countries where consultants hail from.
- **Abstract/executive summary:** A maximum of 300 characters describing what the project is all about and the data collection activities to be or undertaken
- **Key words:** A maximum of five keys words, separated by semi-colons, to identify the project
- **Start Year and Month:** Year and month when the project started or expected to start
- **End Year and Month:** Year and month when the project ended or expected to end