– IMPACT Initiatives –
Data Cleaning Minimum Standards Checklist

Date	13/01/2020
From	HQ Research Design & Data (RDD) Unit Manager
To	IMPACT Country Teams
Subject	Establishing minimum standards for data cleaning & processing at IMPACT

1. Introduction

This memo aims at establishing the **minimum standards that all IMPACT teams (field and HQ) are expected to adhere to in the data cleaning and processing stage of the research cycle.** These standards **must be fulfilled for a dataset to be considered validated and usable for analysis.** If for some reason the standards are anticipated to not be possible to fulfil for a specific research cycle, please flag this to the HQ RDD Unit Manager as early as possible, so a solution can be found.

The minimum standards checklist are aimed to align with A) the IMPACT Data Protection SoPs and B) the forthcoming IMPACT Data Cleaning Guidelines.

The minimum standards checklist includes (1) specific actions that must be taken during the data processing stage, in addition to (2) key documentation that must be shared alongside any data submitted for validation to HQ. The purpose of the key documentation is to enable the RDD Unit to review the data cleaning process in full and ensure the minimum standards have been fulfilled.

2. Key documentation needed for data validation by HQ

To enable the RDD Unit to verify that the minimum standards have been fulfilled during data cleaning, the **following documentation always needs to be shared alongside any dataset submitted for HQ review & validation:**

1. Raw dataset
2. Clean dataset
3. Cleaning sheet/ documentation of cleaning procedure and the types of checks done; for example:
    a. An additional "cleaning" sheet in the Excel database where the "check" columns are included
    b. The R script/ code used to clean and process, and the output generated from this script/ code
4. Cleaning & deletion logs (using the IMPACT Cleaning Logbook template)
5. KOBO/ ODK questionnaire
    a. For all assessments using KOBO/ ODK for data collection, the "audit logging meta question type" should be included during tool design (more here); when possible, data from this should be used to monitor enumerator behaviour (see "Enumerator Metadata" category in Table 1 below)
6. Data Deletion report
7. Sampling verification outputs (see Table 1 below for details)
8. [For qualitative data] A few examples of the raw transcripts and/ or debrief forms used to process and analyse qualitative data

3. How to establish and ensure minimum standards

See Table 1 below.

Table 1: Data cleaning minimum standards checklist

Category	Type of check and relevant action point(s)	Output(s) to be submitted to HQ	When this check should be done	
			During data collection	After data collection
Survey metadata	All records have **unique IDs** or UUIDs. • Action needed: Any duplicates should be deleted and recorded in the data cleaning log.	Deletion log as per the IMPACT Cleaning Logbook template		
Data Protection	All **information that can be used to identify individuals or households is removed** from the dataset. Example of such information: GPS coordinates; Names; Phone numbers; Respondent occupation/ organisation; Information about enumerators / key informants; Respondent gender, age and location; etc. • Action needed: Remove or securely extract all personally identifiable information. All action taken on personally identifiable information should be in line with: (1) IMPACT Data Protection SoPs for Personally Identifiable Information & (2) Indicator risk matrix completed in the ToR (Data Management Plan Annex)	Data Deletion Report as per the template in the IMPACT Data Protection SoPs for Personally Identifiable Information		
Survey metadata	Final dataset is **consistent with intended sampling strategy** i.e.: (1) interview locations/ points and the intended sampling locations/ points are consistent, unless there is a clear rationale (and the limitations of this are well understood);	A clear output (map, table or written summary) outlining the findings from these checks.		

	(2) number of records per stratum match the intended targets per stratum (minus the buffer if this was added to the target to mitigate non-responses); (3) for stratified cluster sampling, interviews with population groups (strata) for a cluster that was sampled based on PPS is within the assigned clusters; (4) there are variables whose values match exactly the strata names in the sampling frame (if applicable); (5) there is a variable whose values match exactly the cluster names in the sampling frame (if applicable). Action needed if any of the above issues are identified: • Any observed diversion should be verified and understood. • Regular tracking during data collection should be done to cross-check the sample collected against the originally intended sample. This could be done either by: (1) preparing an overview map overlaying intended sampling locations with locations where data was collected and (2) maintaining a tracking spreadsheet comparing targets per location per stratum			
Enumerator Metadata	Enumerator **interview speed** (i.e. time taken for the interview/ survey) is reasonable. For most of the assessments implemented by IMPACT, <10 minutes should be a reasonable benchmark. For a more in-depth, comprehensive assessment (e.g. MSNA), the benchmark should be higher (<20 minutes). Ultimately, the benchmark should be based on what is the bare minimum time needed to complete that specific questionnaire, and it should be set by the assessment team during	Deletion log as per the IMPACT Cleaning Logbook template Cleaning sheet/ documentation of cleaning procedure and the types of checks done (for e.g. an additional sheet in the Excel database where the "check" columns are included OR the R script/ code used for cleaning purposes)		

	design and testing, including consideration of multiple survey types (e.g. big/small household, with and without MUAC, different skip logic) • Action needed: If the time taken is lower than expected, additional follow-up should be done to confirm if this is possible.			
Enumerator Metadata	None of the enumerators consistently follow the **shortest questionnaire path OR exact same path** i.e. providing same responses across multiple records. For example, we noticed there is one enumerator (identified by enumerator ID variable) who tends to enter exact same responses across multiple key informants. This seems a bit suspicious and could be an indication of data falsification. The Assessment Officer or the HQ review team might not always have the contextual knowledge to judge whether these are issues indeed or it makes sense that all settlements within the enumeration area have the exact same situation. It is therefore important to follow-up with enumerators to clarify. • Action needed: A clear rationale should be identified for such paths to demonstrate that interviews/ data is not being falsified.	Cleaning log as per the IMPACT Cleaning Logbook template Cleaning sheet/ documentation of cleaning procedure and the types of checks done (for e.g. an additional sheet in the Excel database where the "check" columns are included OR the R script/ code used for cleaning purposes)		
Logical checks	There are **no inexplicable or impossible outliers** i.e. an observation/ a specific data points that lies an abnormal distance from other values in the dataset. For example, if we know the average income in a specific area is around 500 USD/ month, if a household reports an income of 100,000 USD, this could be the result of a data entry error. • Action needed: All outliers should be identified, investigated and corrected as appropriate. • Action needed: It is also important that identified outliers are not automatically assumed to be incorrect and deleted without follow-up. In the example provided above, such high income levels could	Cleaning log as per the IMPACT Cleaning Logbook template Cleaning sheet/ documentation of cleaning procedure and the types of checks done (for e.g. an additional sheet in the Excel database where the "check" columns are included OR the R script/ code used for cleaning purposes)		

	be possible e.g. if the household size is bigger than the average for that area. In other words, sometimes what we consider to be an outlier might not necessarily be one.			
Logical checks	There is **logical coherence between the different responses** within a record. During daily data cleaning at country level, the types of logical inconsistencies to look out for and the action to be taken if such an inconsistency is identified should be clear for everyone working on the cleaning process. During HQ review, it is not possible to identify and check every possible logical inconsistency on every dataset. However, the more obvious potential inconsistencies are checked and flagged. For e.g.: All KIs say that (1) MOST people were unable to access food in the last month and (2) that MOST people did not have access to their usual livelihood activity. However, they also say (3) "Hunger is small, strategies are available to cope" and (4) when there was not enough food, the strategy people used was to "Borrow food from others". The link between these responses from the same KI for the same settlement may be illogical so must be checked. • Action needed: Inconsistencies between questions should be identified, investigated and corrected as appropriate. • Action needed: Follow up questions should be double checked for coherence with top level questions (e.g. reported levels of access to food and use of strategies to cope with a lack of food). • Action needed: Double check that within each variable, all data has the same unit (e.g. number of days or currency in US Dollars) in all rows.	Cleaning log as per the IMPACT Cleaning Logbook template Cleaning sheet/ documentation of cleaning procedure and the types of checks done (for e.g. an additional sheet in the Excel database where the "check" columns are included OR the R script/ code used for cleaning purposes)		

Cleaning log	A **clear, comprehensive cleaning log is maintained** as per the IMPACT Cleaning Logbook template**.** All the different types of data checks done and the follow-up action(s) taken should be evident by looking at the log.
	
	• Action needed: Add exactly one row for each individual data entry that was flagged during the daily data checks.
	
	• Action needed: Conduct a final check that the number of cleaning log entries (= the number of checks) is reasonable given the type of questionnaire and context of data collection.
	
	  ○ For direct management of data collection, suggested benchmark is 5% of total records for which issues were identified, followed up
	
	  ○ For remote management of data collection, suggested benchmark is 10% of total records for which issues were identified, followed up
		Cleaning log as per the IMPACT Cleaning Logbook template
		
		In the unlikely event that no issues were identified during the data cleaning process, a note should still be left in the cleaning log explaining the checks that were done and why no issues were identified despite these checks.		

Cleaning log	A **clear, comprehensive deletion log is maintained** as per the IMPACT Cleaning Logbook template. • Action needed: Add exactly one row for each survey record deleted; rationale behind the deletion should be clear from the log, based on minimum standards established from the outset by the assessment team to determine when records may need to be deleted in their entirety. • Action needed: At the end of data collection, if a high percentage of surveys (>10%) have to be removed because of data quality, no consent or enumerator errors, any biases introduced as a result should be clearly flagged when presenting the findings.	Deletion log as per the IMPACT Cleaning Logbook template		

Data formatting	Dataset is in a **clean, tidy and usable format for purpose of analysis** • Action needed: "Other" responses have been recoded into existing categories or new categories as relevant • Action needed: Missing data fields are left blank or replaced by NA where needed. • Action needed: Within each variable, it is checked that all data has the same unit (e.g. number of days or currency in US Dollars) in all rows. • Action needed: For numeric variables, if for data collection other codes were introduced (ie.999 - not recommended), these are replaced by blank or NA in the final cleaned dataset. • Action needed: Where relevant, the variable within the dataset that should be used for calculating and applying weights should be clearly labelled.	Final, cleaned dataset to be used for analysis		
Data formatting	Dataset is in a **clean, tidy and usable** format for anyone not familiar with the research, with a clear READ_ME sheet • Action needed: All steps outlined in the [forthcoming] Data Cleaning Guidelines for dataset publication should be taken.	Final, cleaned dataset to be published		