IMPACT Minimum Standards Checklist for Quantitative Data Analysis Validation (Probability Samples)

This section aims at summarising all the necessary documents and steps needed to achieve a fast and thorough validation of the analysis of probability samples.

1. Key Documentation Needed for Analysis Validation by HQ

The following list contains the documentation needed by the HQ RDD Unit to validate data analyses performed on probability samples.

1. (Extended) Data Analysis Plan – this can either correspond to the Data Analysis Plan validated at Research Design stage, or, if changes are done to it, and if the analysis is extended with more details, it needs to reflect the steps of the analysis sent for validation, and highlight the differences with respect to the initial Data Analysis Plan. A template can be found in the Probability Sampling Analysis Preparation File.
2. README – a short analysis description summarising all the steps and relevant documents/sheets/scripts for them. It can be sent in your preferred format, but the Probability Sampling Analysis Preparation File can be used as a template.
3. Input clean dataset (after validation, as per Data Cleaning Guidelines)
4. Sampling frame and relevant sampling information (type of sampling, confidence level and margin of error). You can refer to the Probability Sampling Analysis Preparation File as template.
5. All relevant code, scripts, excel sheets and calculations (as per README details)
6. Analysis outputs in easily readable format (xlsx/csv) – as much as possible provide separate outputs for each step of the analysis (e.g. if descriptive statistics and composite indices calculated, provide two separate outputs; if intermediate steps of aggregation/data manipulation, provide them as separate outputs)

2. Preferred Tools

The RDD Unit has capacity to review analyses performed with the different tools listed below, but has preference for the ones highlighted in bold:

a) ***R*** – organisational language, support provided by HQ and country teams, standardised organisation analysis tools available (hypegrammaR)
b) ***Excel*** – suggested tool where no coding capacity for a specific assessment, HQ has templates that can be provided / can assist on more complex operations
c) SPSS
d) Stata
e) Tableau

If you are considering using any other tools apart from the ones above as this is better suited to your analysis needs, please do reach out to HQ RDD Unit first to ensure there will be capacity to review this within the timeframe needed.

3. Summary of Minimum Standards

The table below summarises analysis standards that IMPACT aims to fulfil at the moment or in the near future for the analysis of probability samples. The table is to be read as follows:

- The Standard is an analysis principle common to research in several fields
- The Tools column highlights with which tools from the above list (where relevant) the standard can be achieved, and where there may be need for a check-in with the RDD Unit. Where "different levels of feasibility" is reported, the recommendation to resort to more flexible analysis tools if the team capacity allows, since generally Excel/Tableau will perform more poorly.
- The Requirement column defines those that can be considered minimum standards. All the Required standards need to be accomplished by all analyses for each of the six mentioned topics.

Table 1: Standards for analysis of probability samples

Standard	Tools	Requirement
T1. General		
a) **Dataset** must be **validated** by HQ RDD Unit prior to the analysis, according to IMPACT's Data Cleaning Minimum Standards (checklist in en, fr, guidelines)	N/A	**Required**
b) **ToR/methodology note and (minimal) data analysis plan** (templates here) must be **validated** by HQ RDD Unit before the data analysis starts	N/A	**Required**
T2. Analysis Plan (strictly interconnected with Research Design stage)		
a) **Aggregation** method **reflecting the data analysis plan** should be used (mean, median etc.)	All tools (unweighted aggregations), for weighted medians with Excel/Tableau, contact HQ RDD Unit	**Required**
b) Analysis that does not serve to directly answer a (sub)research question must be minimised. There should always be a **well-defined hypothesis associated with the analysis**	N/A	**Required**
c) Analysis that will not be directly or indirectly reported on must be minimised. **All hypotheses that were tested should also be reported on**	N/A	**Required**
d) **Independent variable(s) must be used only as specified in the data analysis plan** (for exceptions contact HQ RDD Unit)	All tools	**Required**
e) **Comparison** between groups must be made only **as specified in the disaggregation column of the data analysis plan** and as directly relevant to the research question (for exceptions contact HQ RDD Unit)	All tools	**Required**
f) **Comparison between groups that goes beyond the sampling stratification plan** (e.g. gender of Head of Household, which was not included initially as a stratum for the sample design) should be linked to a clear hypotheses and implemented with the relevant test for statistical significance (e.g. chi-square for categorical variables, t-test for continuous variables)[1] prior to reporting these findings	All tools	**Required**

[1] More detailed guidance on which significance to implement and when will be outlined in the (forthcoming) detailed guidelines for Quantitative Analysis for Probability Samples.

in the final information products.[2]		
g) **No exploratory analysis on collected data should be performed.** Exploratory analysis generates hypotheses, and confirmatory analysis quantifies and tests hypotheses. IMPACT quantitative assessments are generally aimed at confirming hypotheses. If planning to do additional exploratory analysis outside of the initial DAP, please contact HQ RDD Unit.	N/A	**Required**
T3. Representativeness & Generalisability		
a) **Unequal sampling probabilities must be corrected with weights** (for example when aggregating strata in stratified probability sampling)	All tools, with different levels of feasibility	**Required**
b) **Weights** used for stratified probability sampling data should be **adjusted** for effective per-indicator sample size where relevant (large number of NAs)	All tools, for more info, contact HQ RDD Unit	Where capacity
c) **Post-stratification weights** should be considered to correct for non-responses	All tools, with different fesibility levels	Where capacity
T4. Uncertainty Estimation		
a) General **certainty** (margin of error and confidence level) **must be calculated, declared and further reported** in methodology/limitation section of the output for both **full sample and subsets** (skip logic). For **cluster sampling**, refer to the relevant MEMO	N/A	**Required**
b) **Calculated certainty should be declared**: Confidence intervals (error bars for graphs) and p-values (hypothesis tests) reported.	All tools, with different feasibility levels	Where capacity
c) **Certainty calculations and hypothesis tests should be adjusted according to the sampling strategy**. Error margins and confidence intervals should be corrected for cluster sampling design effect (see a). Hypothesis tests should make the Rao-Scott adjustments for stratified sampling.	All but excel and Tableau	Where capacity
d) **Hypothesis tests should be adjusted for multiple testing**. If multiple hypothesis tests are performed on the same dataset, they should be corrected using False Discovery Rate procedures or the Bonferroni correction	All but excel and Tableau	Where capacity
T5. Comparability		
a) **Comparisons over time** should only be done on the **same populations of interest,** or on variations of a given population	All tools	**Required**
b) **Comparisons between groups/locations** should only be **within comparable time frames**	All tools	**Required**
T6. Transparency, Reproducibility and Reusability		

[2] In the final information products, tests for statistical significance and related hypotheses should be explained in the methodology section, while results of those tests (i.e. whether differences were found to be statistically significant or not) should be mentioned in the findings section

a) **All steps of the analysis must be well documented, explained and easily understandable.**	All tools	**Required**
b) **Any repetition in the analysis must be avoided**. Ideally each type of logic (formulas, pivot tables, data transformations,...) should be defined in a single place and reused (functions in code)	All tools	**Required**
c) **Separation of concerns should be applied**. The analysis should be as modular as possible, with different sheets/scripts addressing the different steps of the analysis (data preparation, calculations, visualisations)	All tools	**Required**
d) **Code/Formulas should be "data agnostic".** Formulas/calculations must be as unspecific to the data as possible, to be able to reuse it also on different datasets	All tools	**Required**
e) Any piece of **code** should be **packaged in the most self-contained and reader-friendly fashion**: - R: markdowns and projects - python: Jupyter notebooks - SPSS and Stata: inline comments	R/python/SPSS/Stata	Where capacity
f) **All code should be stored on github** for an easier version tracking, both for personal use, validation, and further sharing	R/python	Where capacity