Photo: G. Smith/The Alliance

CGIAR Standing Panel on Impact Assessment

Data Management and Quality Control Norms

Kyle Emerick
Associate Professor, Tufts University & Panel Member, SPIA

5 March 2025

Photo: N. Palmer/The Alliance

CGIAR Standing Panel on Impact Assessment

Data collection quality assurance

SPIA country studies survey models

National statistical agency conducted surveys

SPIA country team’s own surveys

Photo: Tri Saputro/CIFOR

CGIAR Standing Panel on Impact Assessment

Prior experience: Uganda

In Uganda, the DNA fingerprinting protocol was not followed exactly by the statistical agency. There was imperfect compliance from the enumerators. All of them did not collect the DNA samples. As a result, the DNA fingerprinting result is not nationally representative. The situation could not be properly monitored due to COVID-19 restrictions

Lessons learned

Photo: CGIAR

Liaising with the national statistical agency regularly prior to and during the survey to ensure that all the necessary protocols are being maintained

Conducting extensive field visits to assess the actual scenario on the field

Anticipating full range of options for enumerator non-compliance and ask for relevant data to ensure compliance

Photo: Tri Saputro/CIFOR

Standing Panel on Impact Assessment

Prior experience: Bangladesh

We conducted our own survey through a data firm. SPIA designed all SurveyCTO forms. This provided us real time access to the data allowing us to run high frequency checks and conduct audio audits. That being the case, we were able to identify cases where protocol was not followed and re-survey certain households.

Lessons learned

Photo: CGIAR

Ensuring that proper survey management software is used to collect data which allows audio audits

Having a team in place for audio audits to ensure enumerators are properly asking questions. Placing audio audits strategically in places where you think protocol violations could be most likely

Diligently and regularly conducting high frequency checks to ensure data quality. HFC dashboards can be useful

Photo: N. Palmer/The Alliance

CGIAR Standing Panel on Impact Assessment

Data replication after collection

Ensuring replicability of data

Data cleaning

The data cleaning process from the raw data to the final data should be replicable through the usage of do files, R scripts, etc.

Data merging

The merging of various modules should be replicable through the usage of do files, R scripts, etc.

Data analysis

Data analysis results, figures, tables, etc. should be replicable through the usage of do file, R scripts, etc.

Photo: The Alliance

Photo: N. Palmer/The Alliance

Push button replication: Community of practice?

Do the teams think it would be useful to have a github repository to share code among the teams for high frequency checks, data cleaning, merging and analysis?

Photo: N. Palmer/The Alliance

Open Access and Open Data (OA-OD)

Recognizing the need to make outputs findable, accessible, interoperable, and reusable, CGIAR has made a strong commitment to Open Access and Open Data (OA-OD), with all 15 Centers signing on to the Open Access and Data Management Policy in 2013.

CGIAR Standing Panel on Impact Assessment

Open Access for indicative types of information products

Photo: G. Smith/CIAT

- Peer-reviewed versions of scholarly articles reporting research should be deposited in a suitable repository and made Open Access as soon as possible, ideally at the time of publication, and **no later than 6 months from the date of publication**. Authors are free to choose the journal that is most appropriate to their needs. Where an author publishes in a closed access journal, he/she shall self-archive in an Open Access repository a digital version of the final accepted manuscript (the "postprint" version).
- Information products that are not intended for peer-review journals, such as reports, conference papers, policy briefs and working papers, shall be deposited in suitable repositories and made Open Access as soon as possible and in any event **within 3 months of their completion**.
- The full digital version of books and book chapters shall be made Open Access as soon as possible after publication and in any event **within 6 months** either through self-archiving or other suitable publication arrangements.

Open Access for indicative types of information products

Photo: G. Smith/CIAT

- Data (and any relevant data collection and analysis tools) shall, subject to any additional donor requirements, be deposited in a suitable repository and made Open Access as soon as possible and in any event **within 12 months of completion of the data collection or appropriate project milestone, or within 6 months of publication of the information products underpinned by that data, whichever is sooner**. Data deposited shall be prepared in a manner consistent with the aims of this Policy. Existing and future databases shall be made Open Access.
- Complete final digital versions of video and audio outputs, and image collections must be stored appropriately and made Open Access **within 3 months** of their completion.
- Where an information product is software developed internally, the associated source code must be deposited in a free/open software archive upon completion of the software development. Access to such information products may be granted subject to appropriate licences (e.g. Copyleft).
- The metadata of an information product must be deposited in a suitable repository before or on publication of the information product. Where an information product is not deposited in a suitable repository, the deposited metadata must include a link to the information product.

CGIAR Standing Panel on Impact Assessment

Photo: N. Palmer/The Alliance

J-PAL essential checklist for data publication

Folder structure for submission

Photo: G. Smith/CIAT

CGIAR Standing Panel on Impact Assessment

○ Publish a set of files related to a research project by saving them in a clear file and folder structure and then compressing the entire set of folders into e.g., a zip archive. The folder structure might look something like the following:

➢ Main folder

➢ Data

➢ Code

➢ Output

➢ Additional documentation (such as survey instruments, codebooks, etc.)

➢ Readme

Data

Photo: G. Smith/CIAT

- Provide the data in a file format that can be used independently of statistical package choice, such as csv files.
- Ensure that your data is de-identified.
- Include all variables, treatment conditions, and observations collected from the implemented survey instruments (excluding PII)––if feasible and allowed by the data provider (for administrative data).
- If you have multiple datasets, ensure there are ID variables in each dataset that link across datasets
- Ensure the ID variable uniquely identify observations
- Ensure correctly and consistently coded and labeled missing values
- Ensure variables match with the accompanying questionnaire

Code

Photo: G. Smith/CIAT

- Include programs and scripts needed for a push-button replication of all published results
- Make sure code files have headers (including the name of the person who last wrote/edited the code, the date, and the software and version used).
- Make sure code has comments or is self-documenting.
- Remove unnecessary code that creates tables and figures that are not included in the main results or appendix of the paper.
- Remove unnecessary comments and ensure that no identifying information is included in comments
- Execute final versions of the code. Once all the code has been edited to accommodate the changes in the data, run it to ensure it runs without error and that the new results are consistent with those reported in the accompanying published or working paper.

Photo: Tri Saputro/CIFOR

SPIA support areas

Within this domain, which areas do the teams envision the most challenges?

CGIAR Standing Panel on Impact Assessment

Thank you

www.linkedin.com/company/iaes-cgiar/

@cgiarspia

www.iaes.cgiar.org/spia

Independent Advisory and Evaluation Service