DATA COLLECTION AND QUALITY CONTROL

LESSONS FROM HRITF COUNTRIES
RBF studies: General Description and Lessons

- Sample Size & Sample Points (SPs) Quantity
- Teams and Team Composition
- Fieldwork period
- Data Entry
- Fieldwork Centralized control system
Teams and Team Composition

- Teams Quantity: depending on the country: between 6 and 16 teams;

- Team Composition: (1) supervisor; (3-4) interviewers; (1-2) anthropometric experts; (1) field-quality checker
<table>
<thead>
<tr>
<th>Sample Size &amp; SPs Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(Oct 5, 2010)</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Household</th>
<th>Health Facility</th>
<th>Health Worker</th>
<th>Patient Exit Interviews</th>
<th>Community Health Worker</th>
<th>Exposure Period (months between baseline and endline)</th>
<th>Rounds of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5016</td>
<td>2048</td>
<td>1060</td>
<td>3600</td>
<td>TBD</td>
<td>12-24 months</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>152</td>
<td>320</td>
<td>152</td>
<td>300</td>
<td>TBD</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>TBD</td>
<td>800</td>
<td>522</td>
<td>0</td>
<td>TBD</td>
<td>24?</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>TBD</td>
<td>2400</td>
<td>874</td>
<td>500</td>
<td>TBD</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>0</td>
<td>0</td>
<td>500</td>
<td>TBD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Exposure Period (months between baseline and endline): 12-24 months
Rounds of Data: 2
The smaller HH sample case

- SPs = 107
- Typical amount of HHs x SP = 10
- Fieldwork time frame = from Sept 06 to Nov 23, 2009 (totalizing 78 days)

Estimated time in SP = less than 2 days
Tasks to do in 2 days

- Unpack
- Contacting the local authorities;
- (not always, but in some cases) listing operation and household selection;
- Identify and Contact the Households;
- Interview the Households + Health facilities + Health Workers + Exit Interviews;
- Quality Control (QC) (checking, identifying errors and inconsistencies);
- Deal with the inconsistencies (second visit to a HH);
- QC second round(s) (CAFÉ or not);
- If CAFÉ: Data Entry (in one or two afternoons)
- Pack and leave (!)
Lessons and some Inferences

- Regarding the amount of teams and their composition:
  - Fewer teams are more likely to be effectively supervised than a larger number of them (ratio fieldwork period and teams to be visited);
  - The larger the number of questionnaires to be checked (per day), smaller is the chance of performing a good QC;
  - Mobilizing less teams for a longer period is as expensive as mobilizing more teams for a shorter period;
Lessons and some Inferences

- Regarding the Fieldwork design:
  - Short time in the SP: it is more likely to have replaced Households;
  - The larger the number of questionnaires to be checked in a short period of time, smaller the chance of performing a good QC job;
  - Most of errors and inconsistencies are fixed AFTER the fieldwork period (identified post-hoc)
  - Absence of inconsistencies in final databases does NOT guarantee data’s external validity
Lessons and some Inferences

- In simple terms: Time and other possible advantages of a short fieldwork period have *by-design* consequences when the questionnaire is complex:
  - Post fieldwork data cleaning period (so time is not actually gained);
  - Preferring the Internal consistency of the data over the external validity of the data.
  - Yet, these consequences are common to the vast majority of surveys (with no coming-back-to-the-source design).
Propositions

- Think again the Fieldwork design allowing to:
  - Decreasing the probability of replacing the 1st choice household;
  - Do real Quality Control (both in-team as core team control): take advantage of integrated design possibilities and involve the core team in this task.
# Scheduling the team's work in each sample point

<table>
<thead>
<tr>
<th>Day</th>
<th>Team leader (cv: control visit)</th>
<th>Interviewer 1</th>
<th>Interviewer 2</th>
<th>Interviewer 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>Enumerating all the households in the sample point</strong>&lt;br&gt;<strong>Selecting the 9 households to be interviewed</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>HH1: cv 1&lt;br&gt;HH2: cv 1&lt;br&gt;HH3: cv 1&lt;br&gt;HH4: cv 1&lt;br&gt;HH5: cv 1&lt;br&gt;HH6: cv 1&lt;br&gt;HH7: cv 1&lt;br&gt;HH8: cv 1&lt;br&gt;HH9: cv 1</td>
<td><strong>Visit 1</strong>: household roster, sections 2-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>HH1: cv 1&lt;br&gt;HH2: cv 1&lt;br&gt;HH3: cv 1&lt;br&gt;HH4: cv 1&lt;br&gt;HH5: cv 1&lt;br&gt;HH6: cv 1&lt;br&gt;HH7: cv 1&lt;br&gt;HH8: cv 1&lt;br&gt;HH9: cv 1</td>
<td><strong>Visit 2</strong>: sections 4 to 7 (+ fix errors*)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>HH1: cv 1&lt;br&gt;HH2: cv 1&lt;br&gt;HH3: cv 1&lt;br&gt;HH4: cv 1&lt;br&gt;HH5: cv 1&lt;br&gt;HH6: cv 1&lt;br&gt;HH7: cv 1&lt;br&gt;HH8: cv 1&lt;br&gt;HH9: cv 1</td>
<td><strong>Visit 3</strong>: sections 7 to 11 (+ *)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>HH1: cv 2&lt;br&gt;HH2: cv 2&lt;br&gt;HH3: cv 2&lt;br&gt;HH4: cv 2&lt;br&gt;HH5: cv 2&lt;br&gt;HH6: cv 2&lt;br&gt;HH7: cv 2&lt;br&gt;HH8: cv 2&lt;br&gt;HH9: cv 2</td>
<td><strong>Visit 4</strong>: sections 12 to 17 (+ *)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td><strong>Accompanying re-visits to the most problematic cases</strong></td>
<td>Re-visited to fix remaining errors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td><strong>Resting and transferring to next sample point</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Scheduling the team's work in each sample point ensures that all households are accounted for, and any issues are addressed systematically. The team leader oversees the process, and each day’s activities are carefully planned to maximize efficiency and accuracy. Re-visits are essential for ensuring that all data are correct, and the team is prepared for the next stage of their research or monitoring efforts.
Scheduling the team's work in each sample point

By design, the length of the questionnaire is no longer a problem: plan the visits to take no more than 2.5 hours.

By design the inconsistencies and errors are solved by asking again to the source.

Less respondent fatigue and ensuring data is consistent with what the respondent did actually say (and not according to what we may assume the respondent did say or what the attribute should have been according to an imputation system).
Enumerating all the households in the sample point

Selecting the 9 households to be interviewed

Visit 1: household roster, sections 1 to 3

Visit 2: sections 4 to 7 (+ fix errors*)

Visit 3: sections 7 to 11 (+ *)

Visit 4: sections 12 to 17 (+ *)

Accompanying re-visits to the most problematic cases

Re-visiting to fix remaining errors

Resting and transferring to next sample point

INCREASE THE CHANCES OF INTERVIEWING THE 1ST CHOICE HOUSEHOLD

ALLOWS CONTROL VISITS ON A DAILY BASIS

ALLOWS CHECKING ERRORS/INCONSISTENCIES ON A DAILY BASIS

ALLOWS DEALING WITH ERRORS/INCONSISTENCIES ON A DAILY BASIS

RE-THINK THE PRODUCTIVITY CONCEPT: FROM [INTERVIEWS X DAY X INTERVIEWER] TO [HHS X TEAMS X SP]

THIS TASK IS BETTER PERFORMED BY A MACHINE
Controlling the non-sampling error

Integrating the DEP into fieldwork
Controlling the non-sampling error
the role of computer-assisted field edits (CAFÉ)

- Objective: besides the data entry, this system must be able to provide reliable feedback regarding the quality of the data appearing in the questionnaires in a timely fashion (when in the field, on a daily basis).
Controlling the non-sampling error
the role of computer-assisted field edits (CAFÉ)

- Functionalities:
  - Detecting errors, such as a 13-year old person (in the household roster) appears as the father of a 5-year old individual in another section.
  - Detecting inconsistencies such as false skipped fields or unexpected information where it should be blank.
Controlling the non-sampling error
the role of computer-assisted field edits (CAFÉ)

- Functionalities:
  - Detecting odd data, such as 3-year old individual measuring 1 meter high. This information is probably an error, but it could be true.
  - Issuing a list of the detected errors, inconsistencies and odd data relating it to a case (household + individual), interviewer and team.
How to make it work?

- What is on the questionnaire goes to the DB. Then, the DEP should allow out-of-range data (issuing at the same time the corresponding warning message);
- After the interviewers are back (from the households), the task of the DEO begins;
- DEO enters the data and sorts the error & inconsistencies list;
- At this point, both the supervisor and the concerned interviewer join the DEO;
- The supervisor compares the error & inconsistencies list with what appears on the questionnaire;
Controlling the non-sampling error
the role of computer-assisted field edits (CAFÉ)

How to make it work?

- If the error, inconsistency or the odd data is in the list but not in the questionnaire: this is a typo, then it has to be fixed by the DEO.

- If the error, inconsistency or the odd data is in the list and in the questionnaire: this is an interviewer-related data, then it has to be fixed by the interviewer USUALLY BY THE TIME OF THE NEXT VISIT TO THE HOUSEHOLD.

- The fieldwork at the SP ends when there are no more errors, inconsistencies or odd data that may be explained by poor quality fieldwork.
Controlling the non-sampling error
the role of computer-assisted field edits (CAFÉ)

- **Tech-Specs**
  - One laptop computer per team;
  - Computers: current sub-standard laptops are more than capable to manage any DEP;
  - As such laptops prices are inexpensive, get spare batteries (2 x laptop), license for productivity software (word, excel), license for antivirus (ensure covering the fieldwork period);
  - As you may be concerned by the energy supply in isolated areas, think on acquiring solar panels or car batteries.
Controlling the non-sampling error
the role of computer-assisted field edits (CAFÉ)

- External power supply field-tested solution
Sending data to the Central Team (building the DB)

- So far the best solution has been using courier services such as FedEx or similar for sending USB-drives (budget this);
- Why not wireless? Experience shows that researchers are reluctant to host data in public servers; laptop is used for unintended purposes; virus risk; etc.
- When a schedule of visits to the teams is set, data can be gathered by the Core Team member;
Controlling the non-sampling error
the role of computer-assisted field edits (CAFÉ)

- Preparing the team for CAFÉ (also valid if not CAFÉ)
  - Ideally include the DEOs training in the main training: a good supervisor will be as proficient with the questionnaires as with the DEP;
  - As all the team was trained, any member could replace the DEO if needed (and vice-versa);
  - All prior activities must be aligned in order to have an operational DEP version at the time of the training.
  - Test (and refine) the DEP at the same time of the Pilot operation and the following stages.
Controlling the non-sampling error
Considerations concerning CAFÉ

- Field Data Entry conditions are different from those you find in a centralized data entry room:
  - Field: lack of immediate technical support;
  - No permanent control;
  - DEP can be manipulated by non-experts.

DEP must feature functionalities and characteristics making your application user-friendly and robust.
Controlling the non-sampling error
Considerations concerning CAFÉ

- (specially) If CAFÉ: independently of the platform, the Data Entry Operator (DEO) should access the Data Entry Program by using a menu: Avoid the manipulation of the DEP; helping the DEO will make the DEP more robust.
Controlling the non-sampling error
Considerations concerning CAFÉ

- building the DB
  - It is advisable that the file management was a DEP component’s embedded feature (here in the menu).
Controlling the non-sampling error
Considerations concerning CAFÉ

- CAFÉ special considerations:
  - Remember: DEOs will be working independently and not in a fully-supported and controlled DE room.
  - This is a special challenge for the DEP: all preventive measures to avoid duplicated Households or inconsistent geo-codes must be taken.

Valid codes list: for one team at a precise date. It corresponds to the fieldwork schedule and it’s linked to an external file: the Sample Control File.
Controlling the non-sampling error
Considerations concerning CAFÉ

CAFÉ special considerations: the Sample Control File:

- It is a simple file relating the Sample Points to a particular team, and for enhanced control, to a schedule (can be *xls, *txt).

<table>
<thead>
<tr>
<th>PSU (cluster)</th>
<th>Stratum</th>
<th>Province</th>
<th>District</th>
<th>LLG</th>
<th>Ward</th>
<th>CU</th>
<th>Geographic identifiers</th>
<th>Sampling data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>001</td>
<td>10 Southern Metro</td>
<td>04</td>
<td>NCD</td>
<td>1</td>
<td>NCD</td>
<td>01</td>
<td>NCD 80 GEREHU URBAN</td>
<td>41 - 012 OIKA STREET 246/1</td>
</tr>
<tr>
<td>002</td>
<td>10 Southern Metro</td>
<td>04</td>
<td>NCD</td>
<td>1</td>
<td>NCD</td>
<td>01</td>
<td>NCD 80 GEREHU URBAN</td>
<td>41 - 024 DABU STREET</td>
</tr>
<tr>
<td>003</td>
<td>10 Southern Metro</td>
<td>04</td>
<td>NCD</td>
<td>1</td>
<td>NCD</td>
<td>01</td>
<td>NCD 80 GEREHU URBAN</td>
<td>41 - 047 AGOLO DR. 242/10</td>
</tr>
</tbody>
</table>
Controlling the non-sampling error

Considerations concerning CAFÉ

- DEP’s capability for detecting errors, out-of-range values (comparing entered data to international tables or local values), working with a list of valid codes related to the fieldwork schedule, automatic data management, automatic export to STATA, SAS, SPSS formats, etc., is a complex work that takes **TIME and EXPERTISE**.
Controlling the non-sampling error
Considerations concerning CAFÉ

- Thus, it is advisable to involve the expert (local or TA) at the first stages of the questionnaires’ creation:
  - There are fields needed for programming and creating internal controls that will be suggested by the expert;
  - You will count with an operative version of the DEP by the time of the Pilot and Training.
Controlling the non-sampling error
Data management recommendations

- What to do with the questionnaires?

- In the field, after the data is entered, there is no place for the team keeping the completed forms. They must be sent back to the central team for safe and private storage. Budget this activity.

- Protocols for sending, receiving and storing the forms must be implemented (ensure all the workload was sent, received and stored).

- Centralized double data entry is advisable (randomly selected 5% - 7% of the forms). When CAFÉ: immediately after reception. This practice allows counting with trained reserve staff.
Fieldwork Centralized control system

Control Visits

- **Scheduled Fieldwork Supervision** by core team member: pencil control visits with protocol-defined tasks.

- Control interviews are effective because it has been performed and NOT because it is looking for specific data: people may have changed their opinion, consumption patterns, etc.

- What you are looking for is sending the message that you care about what is going on in the field.
Fieldwork Centralized control system
Assessing the Data Quality

Once the DataBases are received (or after the questionnaires are centrally entered): What can the core team do?

- EXPORT THE DATA TO AN EXPLOITABLE FORMAT AND ANALYZE THE QUALITY OF THE DATA on a regular basis:
  - Select a list of (around) 10 variables (indicators);
  - Study their behavior (intra- and inter-teams comparison)
Fieldwork Centralized control system
Assessing the Data Quality

- Select a list of (around) 10 variables (indicators)
- This analysis shows errors that cannot be identified at the data entry stage
  - Identify them among the critical variables for the study (for example, people declaring illnesses in the last 30 days);
  - Other are extremely susceptible to laziness (for example, Household Size, number of eligible women, number of measured children, etc.)
Example of encountered Problems
Studying the Indicators’ Behavior
Example of encountered Problems

With more time in the field and an integrated research design, the core team will have the information to react and take appropriate measures.

You can do more when dealing with this kind of problem instead of only cleaning the database.

Thank you.