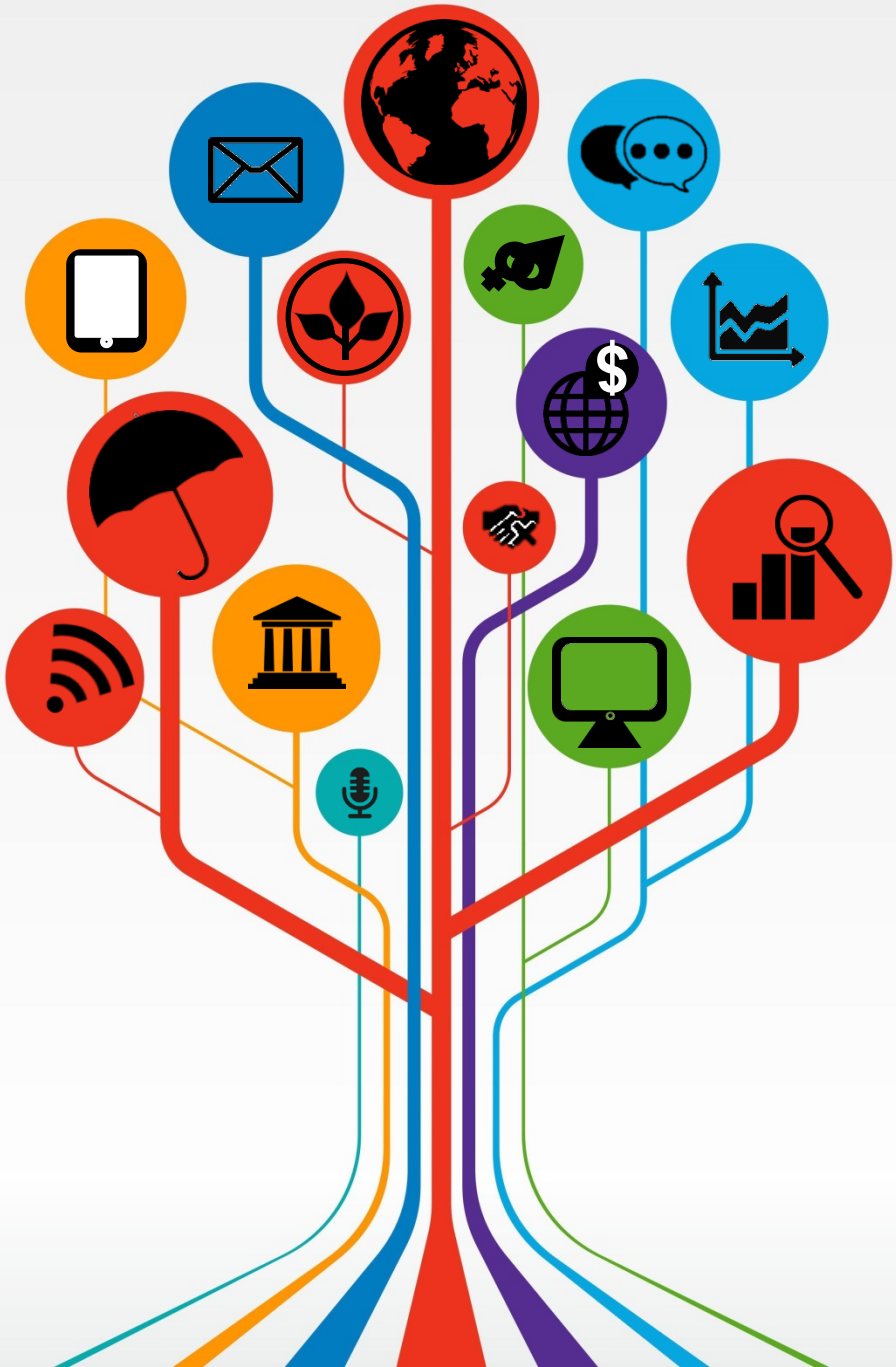
FIELD COORDINATOR WORKSHOP

Manage Successful Impact Evaluations

18 - 22 JUNE 2018
WASHINGTON, DC



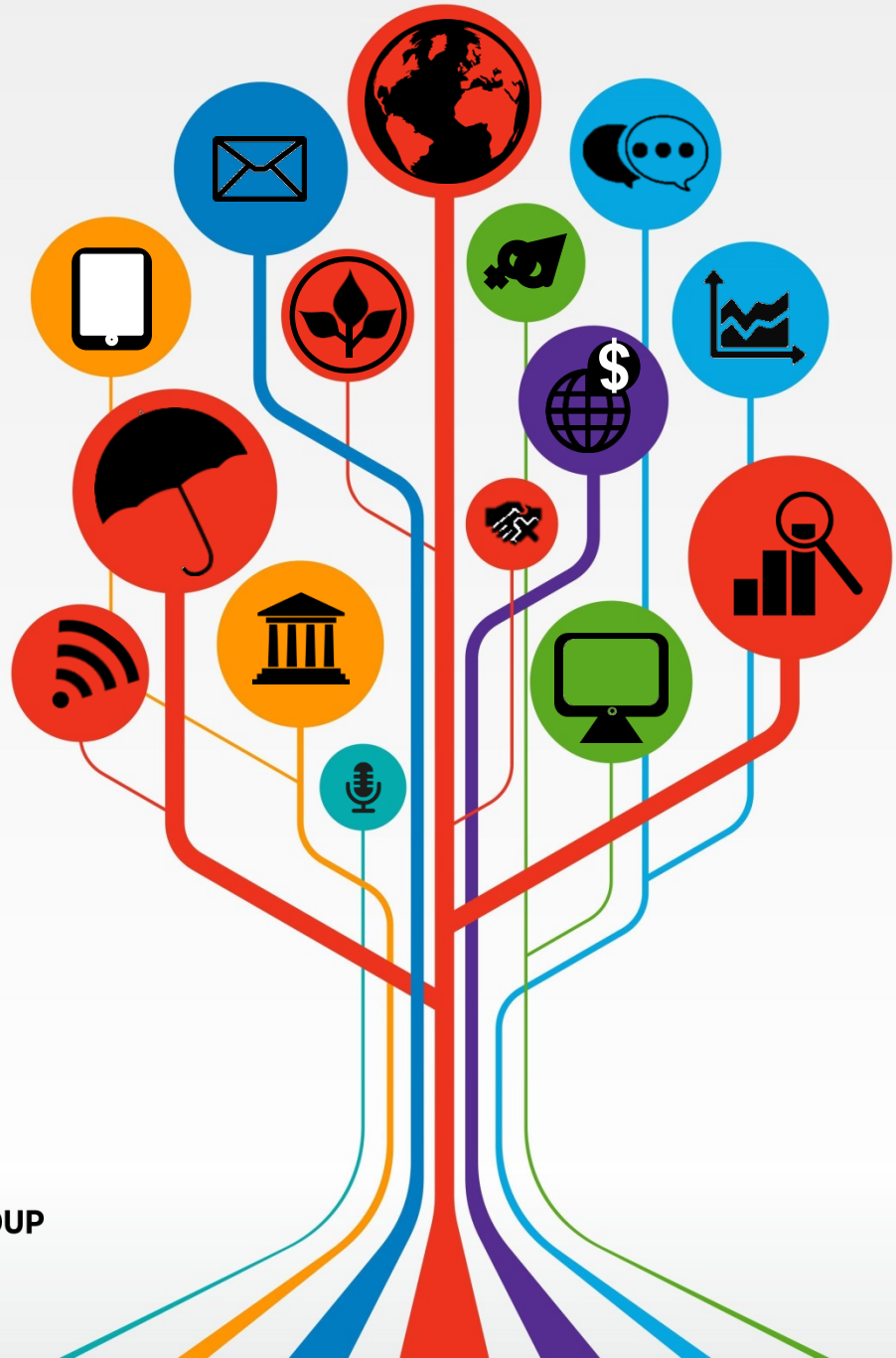
i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP

Data Quality Assurance Plan (2/2)

Laura Costica
Aurelie Rigaud
Steven Glover

19th June 2018



i2i DIME
TRANSFORM DEVELOPMENT

WORLD BANK GROUP

Content

Data Quality Assurance

Before the field

- **Survey programming**
- **Enumerator training**

During the field

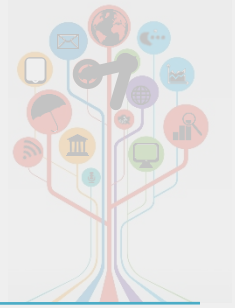
- **Communication and reporting system**
- **Field monitoring**
- **Real time data quality checks**
- **Back checks**

After the field

- **Final field report**

Data related

Data Quality



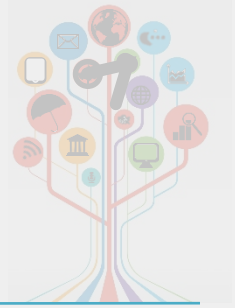
Part 2 -
Data Controls

Data Controls - Content

- Survey Programming
  - Inbuilt programming checks
  - Monitoring checks
  - Testing protocols
- Real Time Data Quality Checks
  - High frequency checks
  - Daily and weekly checks
  - Dealing with errors

Data Quality

Part 2 *Data Controls*



1 - Survey Programming

Survey programming for data quality

- The questionnaire is programmed and uploaded to the SurveyCTO server
- Responses are directly collected into the smartphone/tablet using the SurveyCTO app.
- Surveys are downloaded from the SurveyCTO server to the shared folders of the research team

The CAPI survey is the first place to start to ensure data quality

1 - Inbuilt checks (1/5)

Add inbuilt data quality checks to your programming using hard and soft field constraints:

- Numeric variables (*decimal, integer*)
- String lengths or values (*text*)
- Consistency of selections (*select_one, select_multiple*)
- Other consistency checks (*calculate)*
- Between different fields (*relevance and constraints*)
- Restricting potential (e.g. previously selected) answers (*filter*)

Soft Checks: Warning message, *note*, allowing enumerators to swipe to the next question. The answer is recorded and can be checked later

Hard Checks: Built from field constraints, these checks force enumerators to change the answer to be able to pass. The original answer won't be recorded

Survey Programming

1 - Inbuilt checks (2/5)

No missing values

type	name	label	required
select_one yn	respondent	[3.03] ${hh_name} é um dos entrevistados?	yes
select_one gender	gender	[3.04] Qual é o gênero de ${hh_name}?	yes
integer	age	[3.05] Qual é a idade de ${hh_name}?	yes

CAREFUL: (almost always) never set *note* fields as required!

- Set ALL answers as required to avoid missing values.
- Enumerators can't swipe past if they don't answer a question

Survey Programming

1 - Inbuilt checks (3/5)

Hard checks → Constraints on responses

type	name	label	constraint
integer	plottime	[6.05] Quanto tempo leva para chegar à machamba ${plot_id} da sua casa?	.=-777 or (.>=0 and .<500)
integer	plotyear	[6.06] Em que ano obteve a machamba ${plot_id}?	.=-777 or .=-888 or (.>=1950 and .<=2018)
select_multiple hh	plotresp	[6.07] Quem faz as principais decisões sobre as actividades agrícolas na machamba ${plot_id}?	if(selected(.,-7), count-selected(.)=1, count-selected(.) >=1)

- Responses must satisfy the constraint expression before the survey allows you to pass
- Also can use for select_one and select_multiple types (here we don't allow enumerator to select 'refuse' AND select a household member)

-777 and -7: refuse to answer
-888: don't know

1 - Inbuilt checks (4/5)

Soft checks → Checks using relevance and note fields

type	name	label	relevance
integer	d16	d16. How many trees did you plant over the last 12 months?	
integer	d17	d17. Out of the trees you planted how many trees are still alive?	
note	check1	Be careful, the number of trees alive is higher than the number of trees planted	${d17}>${d16}

- Indicate to enumerators something is inconsistent using a note field as a warning
- Survey allows enumerator to pass through
- Can display the previously entered values for them to check over them
- Use HTML formatting on the label to get their attention:
  - **Bold**: text here
  - Color: text here

Construct the relevance check with previous answers

Survey Programming

1 - Inbuilt checks (5/5)

Using preload data to verify

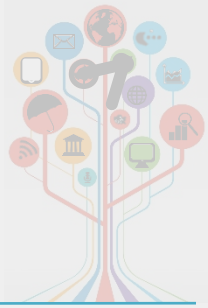
type	name	Label	require	constraint	calculation
text	respondent_id	01- Enter the respondent ID	yes	string-length(.)=5	
text	respondent_idconf	02- Please, re enter the respondent ID as confirmation:	yes	.=${respondent_id}	
calculate	respondent_name	Pull the name of the respondent			pulldata('res
select_one yesno	resp_confirm	03- Do you confirm the respondent is ${respondent_name}?	yes	.=1	

calculation
d}
concat(${loc_id},${respondent_id})
pulldata('respondentname', 'Fullname', 'farmer_id', ${respondent_id})

In the file *respondentname*, the name is pulled from the *Fullname* variable where *farmer_id* is equal to *respondent_id*

- Check HH member names, number of plots, etc.

2 - Monitoring checks



- Other options using SurveyCTO:
  - Random audio auditing (need approval of the respondent, TTL, survey firms…)
  - Text audits (record the time spent on each question)
  - Duration calculation and speed limits to flag forms completed too quickly
  - GPS and options to review surveys collected by geographical location
- Best practice and examples on the DIME Wiki!

3 - Testing (1/3)

Intense testing of the form programming:

Who should test?	How to test?
Survey firm staff, FCs (ask those from other projects), RAs, even PIs!: • Ensure understanding of the instrument before the training • Provides test surveys for HFCs and backchecks coding	Should cover all the combinations of questionnaire structure based on the intervention(s) being tested: • Check skip patterns work correctly • Check *other specify, don't know* work correctly • Check required fields, preload data, calculations, tricky coding all work correctly • Use *note* fields to show stored values at important points

3 - Testing (2/3)

- **How to test?**
  - If survey instrument differs based on the intervention. Example:
    - one form (or set of questions) for control respondents
    - one form (or set of questions) for treated respondents

    2 surveys
  - Cover all response possibilities for each intervention:
    - Answer *yes* and *other specify* all the time
    - Answer *no* all the time
    - Answer *don't know/refuse to answer* all the time

    3 possibilities
  - In this example:
    - At least 6 test surveys should be filled
    - But even so, as many times as possible!

2 x 3 = 6 test surveys

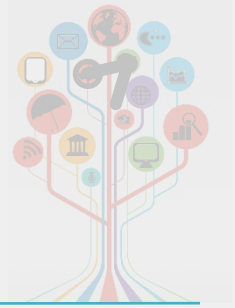
3 - Testing (3/3)

The questionnaire is not fully tested before the data is downloaded and successfully imported in Stata!

- Stata has some restrictions that SCTO doesn't have! Importing to Stata is an important step to anticipate before the field starts
- Use the test dataset:
  - To get familiar with the downloading process and organize your survey data flow
    - Run and edit the import do file
    - Run and edit HFC and cleaning do file
  - To check the variable labels and edit if necessary
    - Labels chosen in the programming can be different than in the dataset (variable created in loops)
  - To edit your programming (add more hard and soft checks…)
- **This is not piloting! All of this should be done before piloting**

Data Quality

Part 2 *Data controls*



2 - Real time data quality checks

High Frequency Check

Definition (1/3)

- HFC : What are high-frequency checks?
  - Different type of checks of the survey data collected
  - Should be run on a daily basis during the field
  - Can't be implemented after the data collection but..
  - Can be completed on regular basis
- How and when to write the code?
  - Create your HFC BEFORE the data collection
  - Use a dataset to write the code (pilot data, test data…)
  - Brainstorm using your questionnaire, identify the questions to checks, the indicators to calculate

High Frequency Check

Definition (2/3)

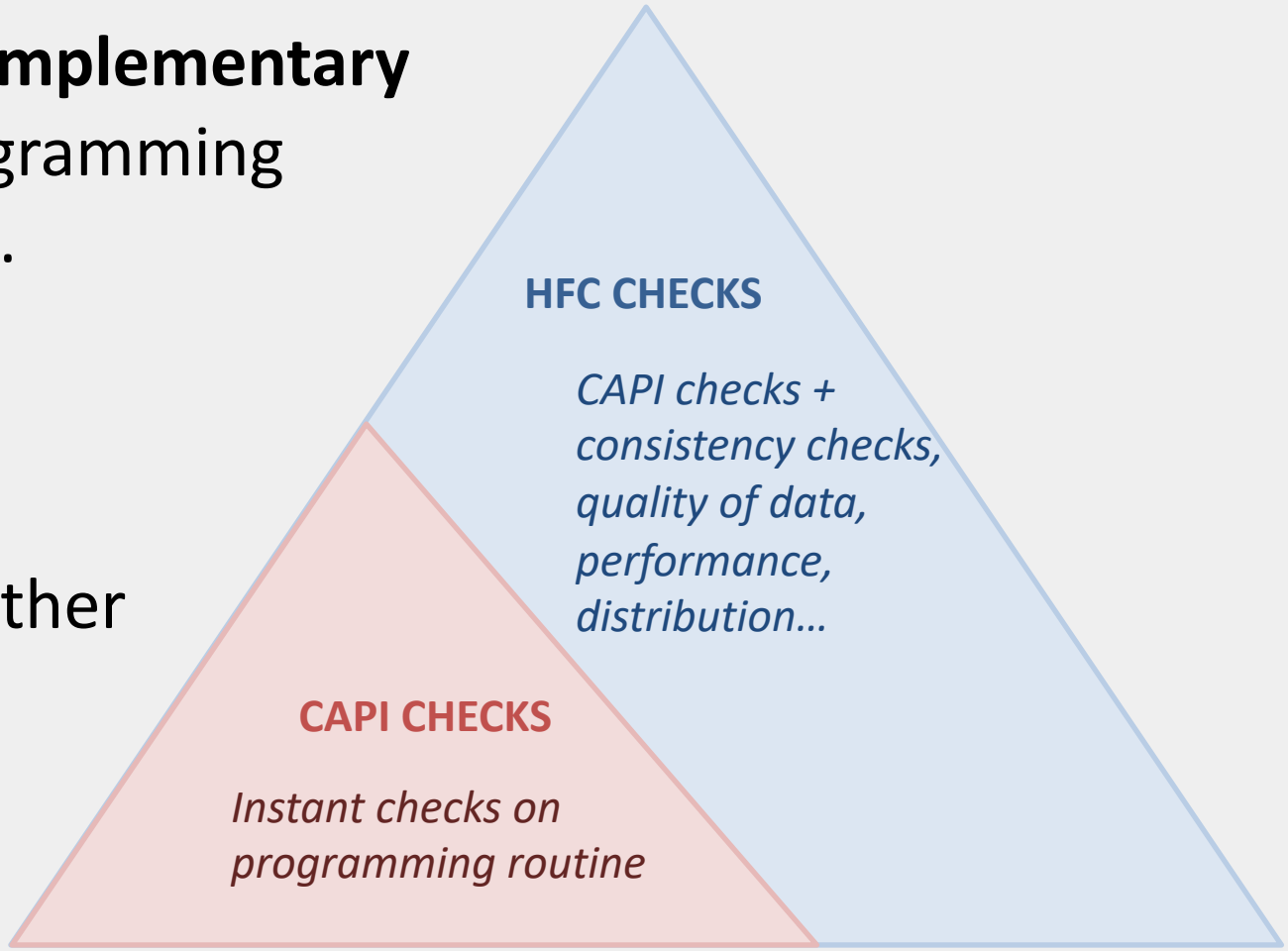
- Some checks should be run daily, others weekly
- HFC provide information about
  - ✓ The quality of the programming
  - ✓ The quality of the data
  - ✓ Enumerator performance and survey progress
  - ✓ Distribution, Data flow,, trends across survey
- Run your HFC each time you download data

High Frequency Check

Definition (3/3)

- HFCs are **complementary** to your programming checks but…

- HFC goes further



High Frequency Check

Daily checks (1/2)

Unique IDs:

Unique ID variables: Respondent code, starting date and time, survey code…

- Are unique ID (and all the other variables that should be unique) actually unique?
- Checks that a form is consistent with other records for its unique ID

Routine and logic checks

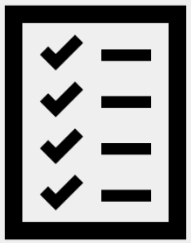
- All the interview should be completed
- Double check key skip patterns
- Double check important hard checks
- Check that no variables have only missing values, where missing indicates a skip

High Frequency Check

Daily checks (2/2)

Date checks:

- Check start & end date of interview are the same
- Start date and time < end date and time
- Check duration



Tracking checks:

- Check the downloaded surveys are consistent with the original tracking list
- Check the surveys sent were all received and downloaded

Consistency checks:

- Logic checks not implemented in the programming

High Frequency Check

Weekly checks (1/2)

Distribution checks:

- % of missing values for each variable
- % of "*don't know*", "*refuse to answer*" for each variable
- Check no variable has only a single distinct value
- Check outliers and extreme values
- Review other specify to ensure consistency of answers entered
- Check the range limits programmed in the form have not been reached

High Frequency Check

Weekly checks (2/2)

Enumerator performances

- ALWAYS review enumerator comments!
- Check number of survey completed by enumerators
- Check average survey duration per enumerators
- Check % of *don't know, zero, no, refuse to answer* per enumerators
- Check enumerator responses on key variables

High Frequency Check

Dealing with errors

- Example of correction sheet to share with field teams
  - Indicate useful information for each problem found (respondent ID, surveyors, date, location, variables…)
  - Explain clearly what is the problem and how to fix it
  - Share general instructions and recommendations
  - Never forget to congratulate!

SLWMP- baseline

Correction Shee Data received before the 6th of March – 3:55pm

Total data received: 376

Total new data checked: 213

ISSUE THAT NEED TO BE CORRE

l numbe	District	Communit	Respon dent ID	Contact	surveyor ID	Surveyo r name	Question/issue	Original Response	What to do?	Correction	Tick if correcti on is made	Comments
1	SW	Gbal	221SW2 501	xxxxxx	C4	xxx	d13_2o: other specify for d13_2 What are the crops you plant in between trees? C0:: What are the crops you cultivates?	d13_2o: "'Soyabean"	Could you confirm the farmer cultivate Soyabean in between trees? Soyabean wasn't selected in C0. Confirm the crops cultivated in C0			
2	SE	Sumboro	011SE09 06	xxxxxx	B2	xxx	D0: inputs used on the land	D0: number of inputs selected: 2 (chemical fertilizer and herbicides)	2 inputs selected by the respondent but questions were answered for 3 inputs. Confirm how many inputs the farmer use? What are they? t? If it is 3 inputs, kindly review the answers for the third input.			
4	SW	Nyentie	024SW18	xxxxxx	C4	xxx	d33_o: other specify for sources of inputs	other specify : "karal"	What does KARAL? Could you clarify whether the respondent was part of another agriculture programme? If yes, a16, a17 and a18 should not be missing. Call the respondent if necessary			

REMARKS / CAREFUL FOR NEXT TIME (all the staff should read this part)

District	Commu	survey or ID	Question/issue	Original Response	FOR NEXT TIME			
SW	Zini	C3 + C6	: c30_o other specify for "What is the main reason that your household cultivates ${crop_label}?"	Consumption and sales / to s	Please select the right option choice. Consumption and/or sale are in the option list!			
WA	Tambiliip	D3	td1_o, td2_o, td3_o otherspecify	Surveyor selected option "other" and entered -99	No need to select other if the respondent doesn't know, enter directly -99			
WM	Sazona	H4	R1: random walk question	The surveyor filled "no" for the question "does the respondent comes from the random walk".	This is a control community so the answer must be "yes" . Read carefully the question.			

TO TELL TO THE WHOLE TEAM (all the staff should read this part)

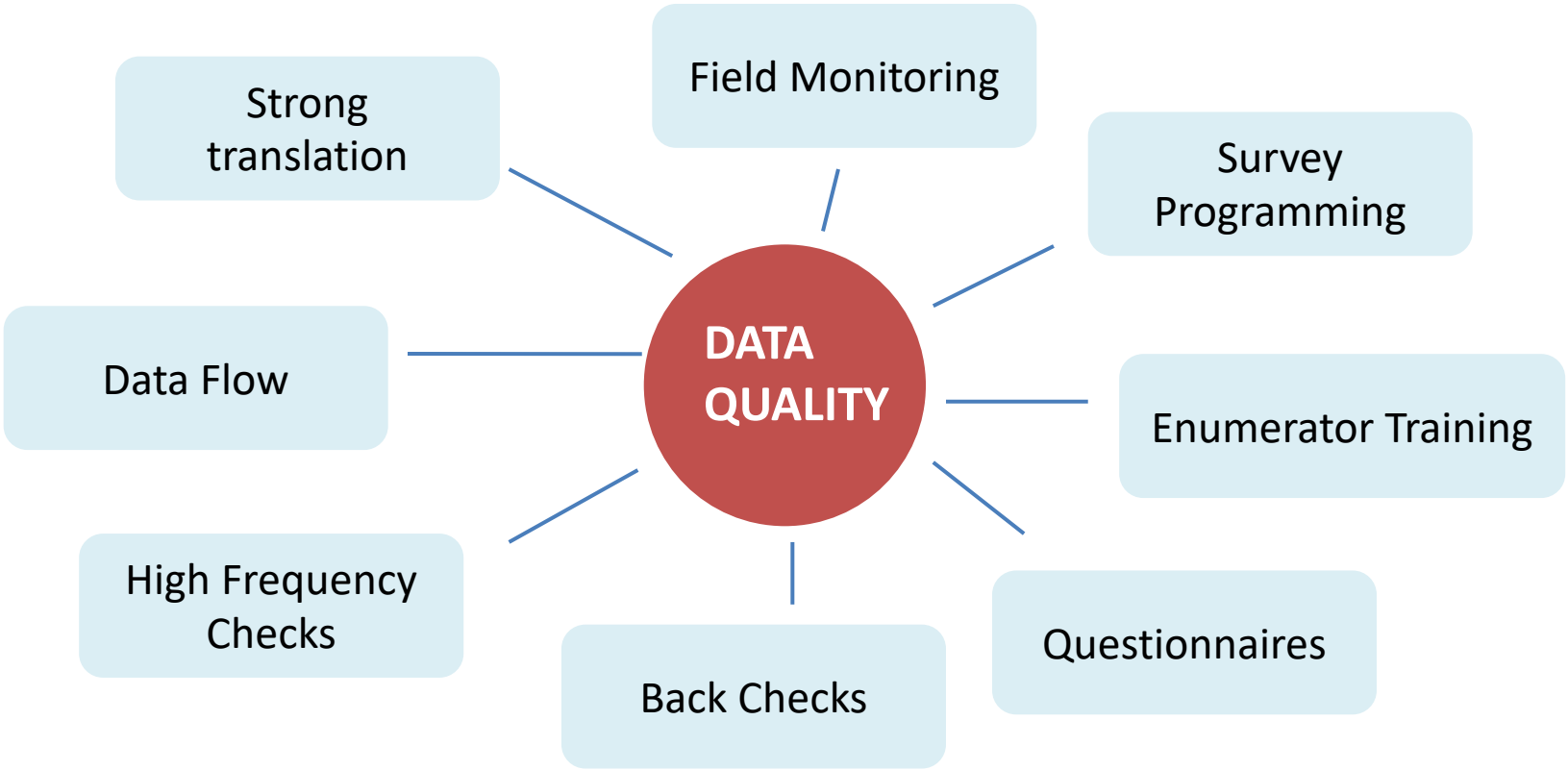
1/ f1. What is the total amount of savings of this household? Please don't fill -99. there are too many "don't know".
Take the time to explain the surveyors what saving means (money kept at home or at the bank,...), enter an estimation
2/ g5. How did you hear about this technology? " learnt from my parents", "it is a tradition", all this is captured in the option choice "family and friends". Do not select other specify ne
3/ Comments: Enter useful comment regarding programming, respondent's understanding of questions. Do not enter: "no comments".

CONGRATULATION

1/ Surveyors C5 and D2 precised in the comment that they entered a wrong number of people in the household and that the number of time questions were repeated was not consiste
2/ Surveyors E2 indicated in the comment the respondent didn't receive all the inputs from the project while he is in the treatment group
3/ No mistake anymore with questions f5_2

Conclusion

Data quality should be considered at every level of the data collection process



Thank you!

Contact us!

arigaud@worldbank.org
lcostica@worldbank.org
sglover1@worldbank.org