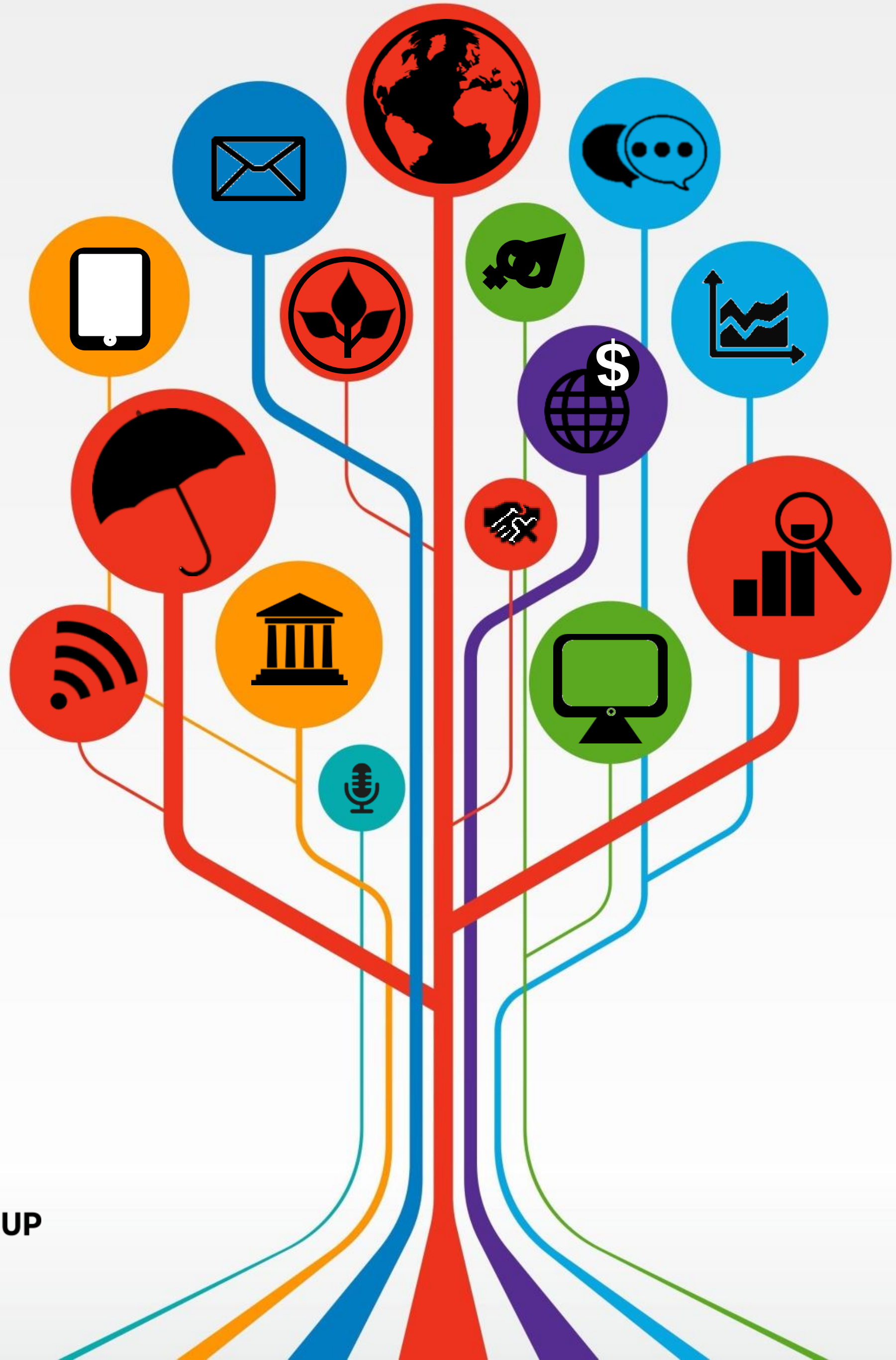
Data Quality Assurance

Steven Glover

Margherita Fornasari

Wednesday 12 June 2019



i2i DIME

TRANSFORM DEVELOPMENT

WORLD BANK GROUP

Data Quality

"The quality of the data we collect plays a key role in driving the quality of our decision-making"

Christopher Robert, *SurveyCTO*

What is quality data?

Data Quality

Think of everything that might go wrong …

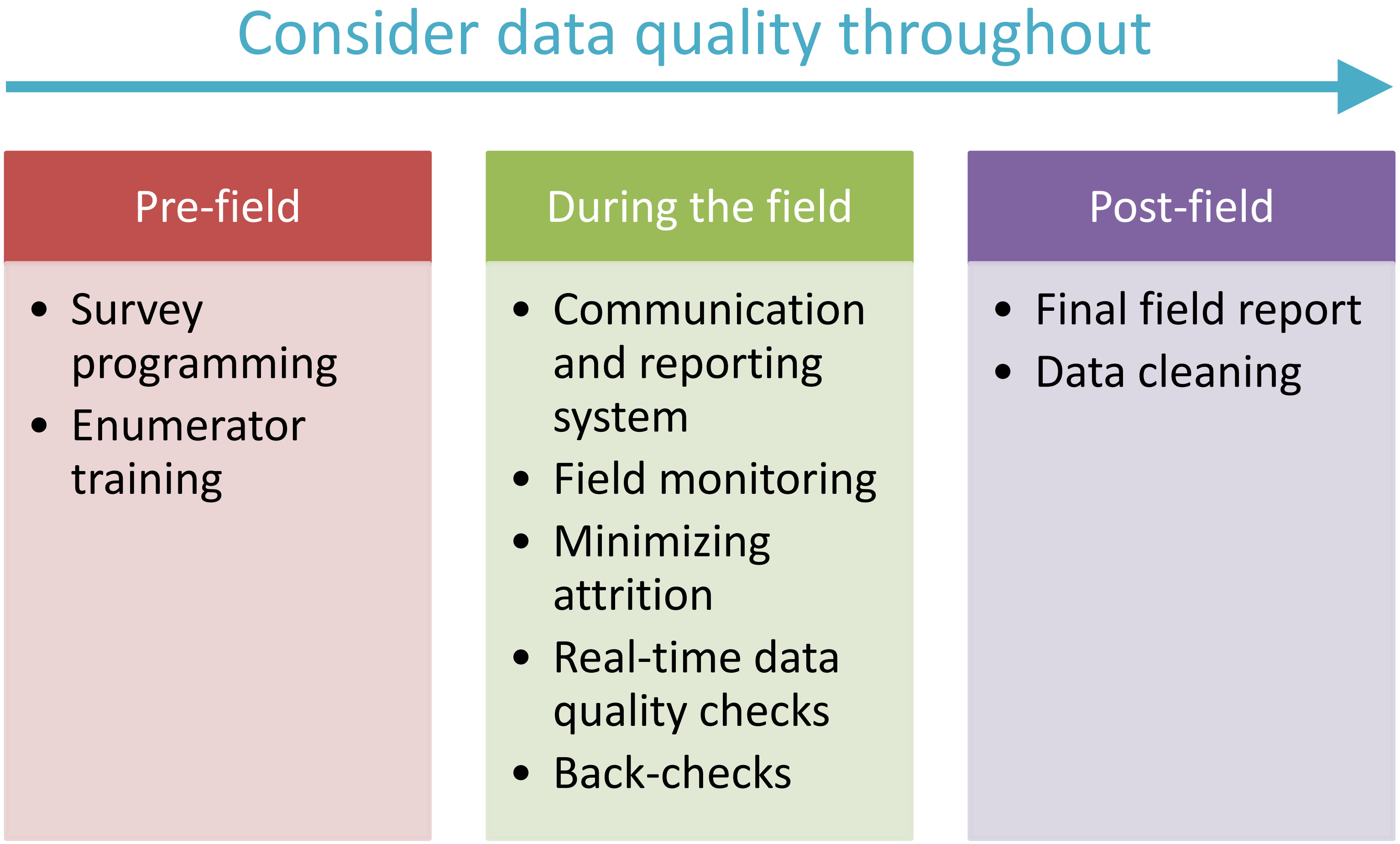
Data Quality

Think of everything that might go wrong …

- Broken tablet or empty battery
- Incorrect recording of answers
- Answers entirely made up
- Sample attrition
- Programming bugs
- No phone signal
- Respondent doesn’t understand the question
- Wrong person interviewed

… and make a plan for pre-empting it

Content



Consider data quality throughout

Pre-field

- Survey programming
- Enumerator training

During the field

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

Post-field

- Final field report
- Data cleaning

Part 1: Pre-field

Pre-field

- Survey programming
- Enumerator training

During the field

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

Post-field

- Final field report
- Data cleaning

Survey Programming

A CAPI survey is the first place to start to ensure data quality:

- Responses collected directly on tablet/phone and able to be sent immediately for analysis
- One-stop-shop – GPS capture, photos, audio…
- User-friendly interface
- Numerous measures to promote data quality:
  - Don't allow missing values
  - Preload data to verify respondent

Survey Programming - Responses

- Inbuilt data quality checks can be used to prevent incorrect information (typos, misunderstandings, etc.) from being entered in the survey
- Piloting and baseline surveys for guidance on limits

Hard checks

- Flag if response is **impossible** – must satisfy a condition
- Do not allow enumerator to continue if answer is flagged
- E.g. household member age is >150 years, negative number of plots

Soft checks

- Flag if response is **implausible**
- Prompt enumerator to verify response if the answer is flagged
- Answer recorded and can be checked later
- E.g. income > $1,000,000,

Survey Programming - Monitoring

Use instrument to verify that survey is being performed as intended:

- Random audio auditing (needs respondent approval)
- Text audits (time spent on each question)
- Duration calculation and speed limits
- GPS location
- Device sensor meta-data

Survey Programming - Testing

Intense testing of the form programming

How to test?

- Check skip patterns work correctly
- Spellings!
- Check *'other specify'* and *don't know* options
- Check hard and soft checks specification
- Check all fields need an answer, preloaded data loads correctly, calculations, and all tricky coding all work as intended
- Use temporary fields for testing that display stored values at important points

Who should test?

- Field coordinators are ultimately responsible
- Survey firm staff, FCs from other projects, RAs, even PIs!
  - Ensure survey firm understanding of survey before training
  - Provides test surveys for data quality checks coding

ietestform – Stata command (part of *iefieldkit*) developed by DIME Analytics that tests SurveyCTO forms for data quality best practices

Survey Programming - Testing

How many times?

- Portions of the survey instrument that differ based on treatment status (or variations of).
  - Treatment questions
  - Control questions

2 levels

- Cover all response possibilities:
  - 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 2 x 3 = 6 test surveys

Survey Programming - Testing

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

- Check that variable names < 32 characters *after* export
- Check that variable names make sense and are consistent
- Use the test dataset to:
  - Get familiar with the downloading process and organize work flow with RAs
    - Run and edit import do file
    - Run and edit HFC and cleaning do file
  - Check variable labels and edit if necessary (esp. from repeat groups)
  - Update your survey form programming (add/edit more hard and soft checks)

THIS IS NOT PILOTING! All this should be done beforehand

Part 2: During the Field

Pre-field

- Survey programming
- Enumerator training

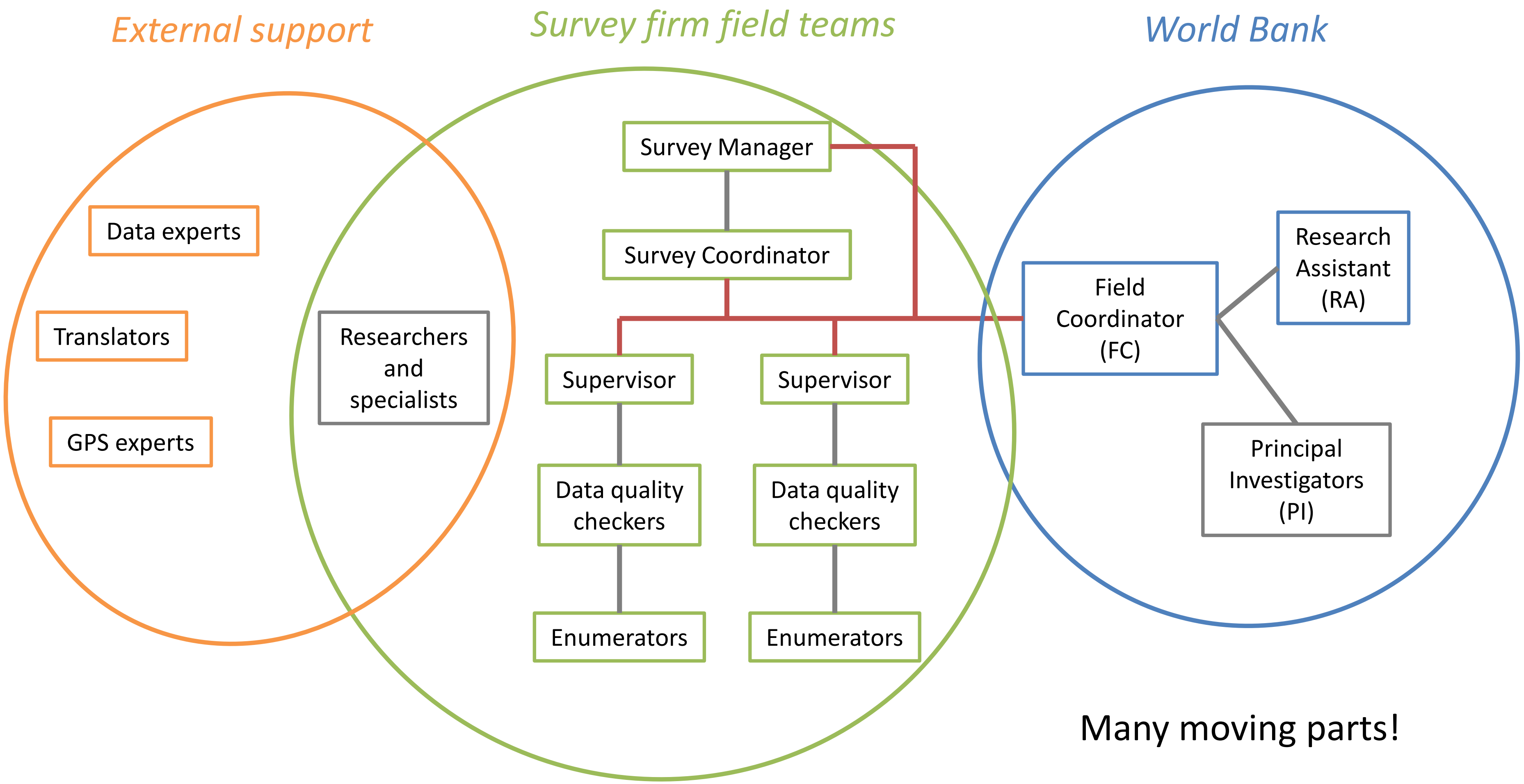
During the field

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

Post-field

- Final field report
- Data cleaning

Survey Team Structure



Many moving parts!

Communication with Field Teams

- Set up a good system for reporting, giving feedback, sample replacements, and responding to queries
  - Create *WhatsApp/Slack* group for key personnel and with each supervisor
  - Shared folder for survey reporting/feedback documentation
- Ensure that each aspect is well understood and practiced before field work starts
- Each field team meets at the end of each day for feedback, sharing experiences – fundamental that messages get to the enumerators

Remote Areas

Surveys in remote areas can present communication issues:

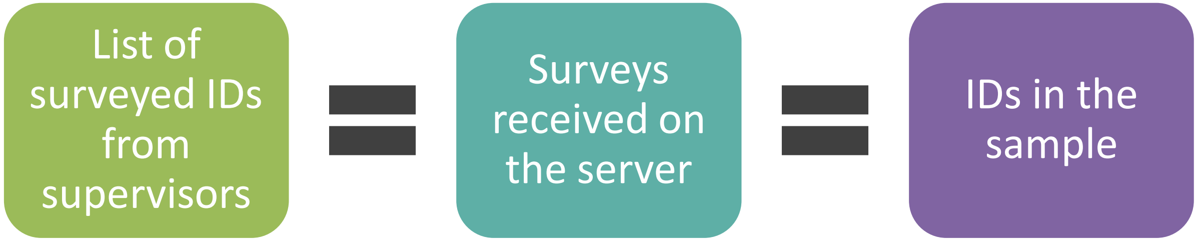
- Cannot send survey forms every day
- Receiving and sending data quality check feedback
- Charging tablets
- Phone network

What can you do?

- Train the supervisors well – many problems don’t need immediate FC interaction to resolve
- Establish protocols for remote areas, e.g. pre-supplied list of replacement households
- Set maximum period before having to check in again or sync tablets (not > 48 hours)
- Ensure field teams equipped with external power banks

Field Monitoring

- Fundamental that surveys conducted match the sample, ensure this through a good reporting system that checks:



List of surveyed IDs from supervisors = Surveys received on the server = IDs in the sample

- Reconcile registers of completed surveys with field teams **every day**. Have a list always accessible to the supervisors (*dropbox / google sheets*)
- Missing cases can create major issues with the survey firm
- Prepare a separate logbook for internal use – this also highlights data correction issues and other notes to understand the survey data

Minimizing Attrition

- Attrition: unmeasured outcomes
- When tracking respondents in a follow-up survey
  - Over 5% frowned upon by peer reviewers
  - Different attrition rate between treatment groups may give biased results!

Issue	Solution
Moved away	- Have a ‘mop-up’ plan, including the budget for it
Cannot be located	- Record identifying info during baseline - Use GPS coordinates to find baseline location - Ask nearby study participants and neighbors
Refuses to participate (or answer certain questions)	- Consent form to put respondent at ease - Survey design - Gift?

Data Quality Checks

High frequency checks (HFCs)

- Run on a **daily basis** for **ALL surveys**
- Check for:
  - ○ Consistency of responses (greater complexity than in programmed survey form)
  - ○ Outlier values
  - ○ Programming checks
  - ○ Enumerator checks
  - ○ Unique IDs, duplicates, dates
- Set up robust and *realistic* system for addressing issues with field teams

vs.

Backchecks

- Revisit households to perform short survey (10-15 mins)
- **10-20% of the sample**, random, frontloaded, for all enumerators
- Check for:
  - Right person interviewed
  - Identify fraud / time-saving
  - If enumerator is recording responses correctly
- Decide on acceptable thresholds and put in place plan to deal with issues

Data Quality Check Workflow

- Produce reports (excel with an issue per row) for both HFCs and backcheck inconsistencies.
- Be clear about what is required by survey firm to deal with the issue
  - Verify value? Redo module? Redo interview?
  - Include info on question number
- Avoid having to go back-and-forth over a single data point, especially if each time requires a trip to talk to respondent

HFC Considerations

- **Always** ask if enumerator can explain the flag – it's not necessarily incorrect
- If multiple errors of same type – stop and re-train enumerators
- **Do not be too ambitious** – identifying and eliminating all mistakes is impossible. Be smart on how to prioritize how you spend your time to get the best data possible.
- Get correct responses for key variables if an error was made, e.g. income, production
- See '*Hands-on Session*' on HFCs for practical guidance on implementation.

Backcheck Considerations

Select questions with both enumerator and questionnaire issues in mind in order to understand the discrepancy:

- Straightforward questions where we expect no variation
  - Number of family members, number of plots
- Questions we expect capable enumerators to get right
  - Questions were perhaps a calculation or estimation is required or on sensitive issues
- Questions we expect to be difficult, see if were correctly interpreted

Timely feedback system to deal with errors:

- Set realistic expectations with actionable steps in each case

Part 3: Post-field

Pre-field

- Survey programming
- Enumerator training

During the field

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

Post-field

- Final field report
- Data cleaning

Final Field Report

- Survey firm usually produces a Final Field Report following the data collection
- Can contribute to data quality when trying to understand the dataset. Qualitative reporting can be used to:
  - Provide information about aspects that could not be captured by the survey instrument:
    - Respondents understanding of certain questions
    - Limited option choices for specific questions
    - Enumerator feedback on understanding of the questionnaire, or fidelity of the responses
  - Share information about community size and structure (sample weights, sample frame)
  - Advice on follow-up survey structure and logistics

Data Cleaning

- Cleaning data is a **key phase** between data collection and analysis
  - Even the best programmed survey requires a little bit of data preparation
- **Decisions on values and data points** are made during this phase
  - Always record changes and ensure replicability
- **Main goals** to keep in mind while cleaning:
  - Identify and clean values that can invalidate or bias the results
    - Outliers
    - Inconsistent values
  - Creating clear, self-explanatory, and informative datasets
    - Variables and values labels
    - Extended missing values approach
    - Reduce the number of string variables
  - Improve the quality of future data collection
    - Data cleaning as a learning process to identify and solve mistakes for future survey design and programming