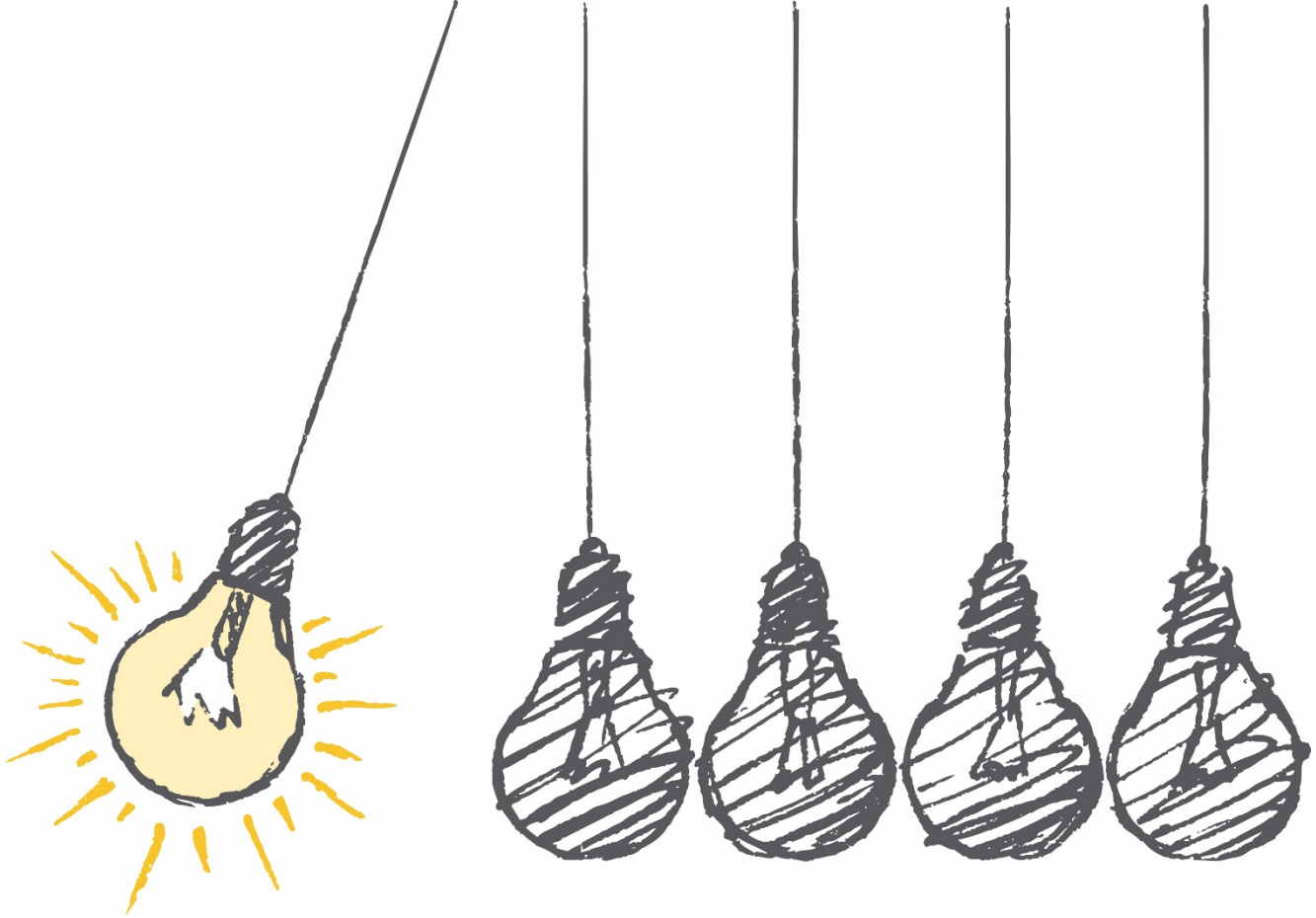
FC Continuing Education Series

Data Quality Assurance During Data Collection



Prepared by **DIME Analytics**
dimeanalytics@worldbank.org

Presented by **Roshni Khincha**
rkhincha@worldbank.org

WORLD BANK GROUP

UKaid
from the British people

Norad

Data Quality

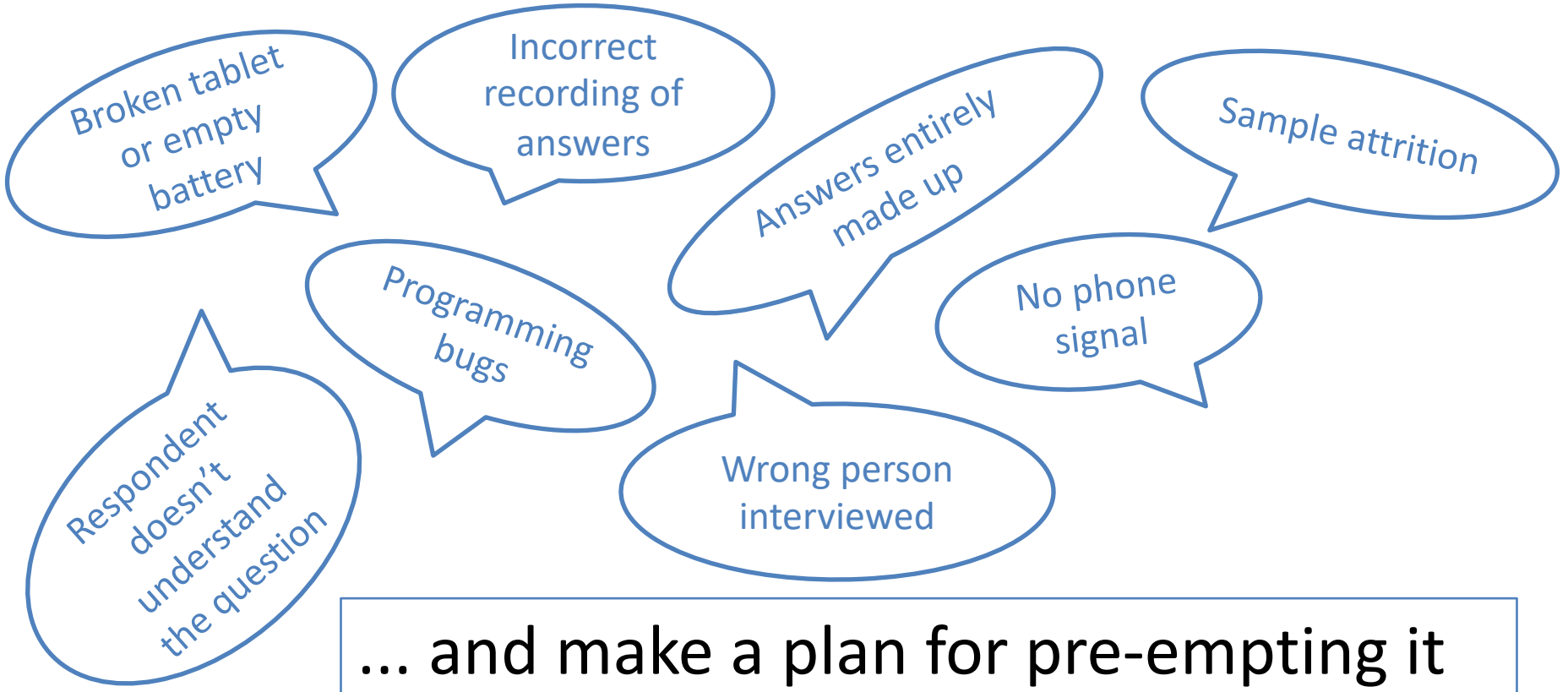
"The quality of the data we collect plays a key role in driving the quality of our decision-making"

Christopher Robert, *SurveyCTO*

What is quality data?

Data Quality

Think of everything that might go wrong…



… and make a plan for pre-empting it

Data Quality for Primary Data Collection

Consider data quality throughout

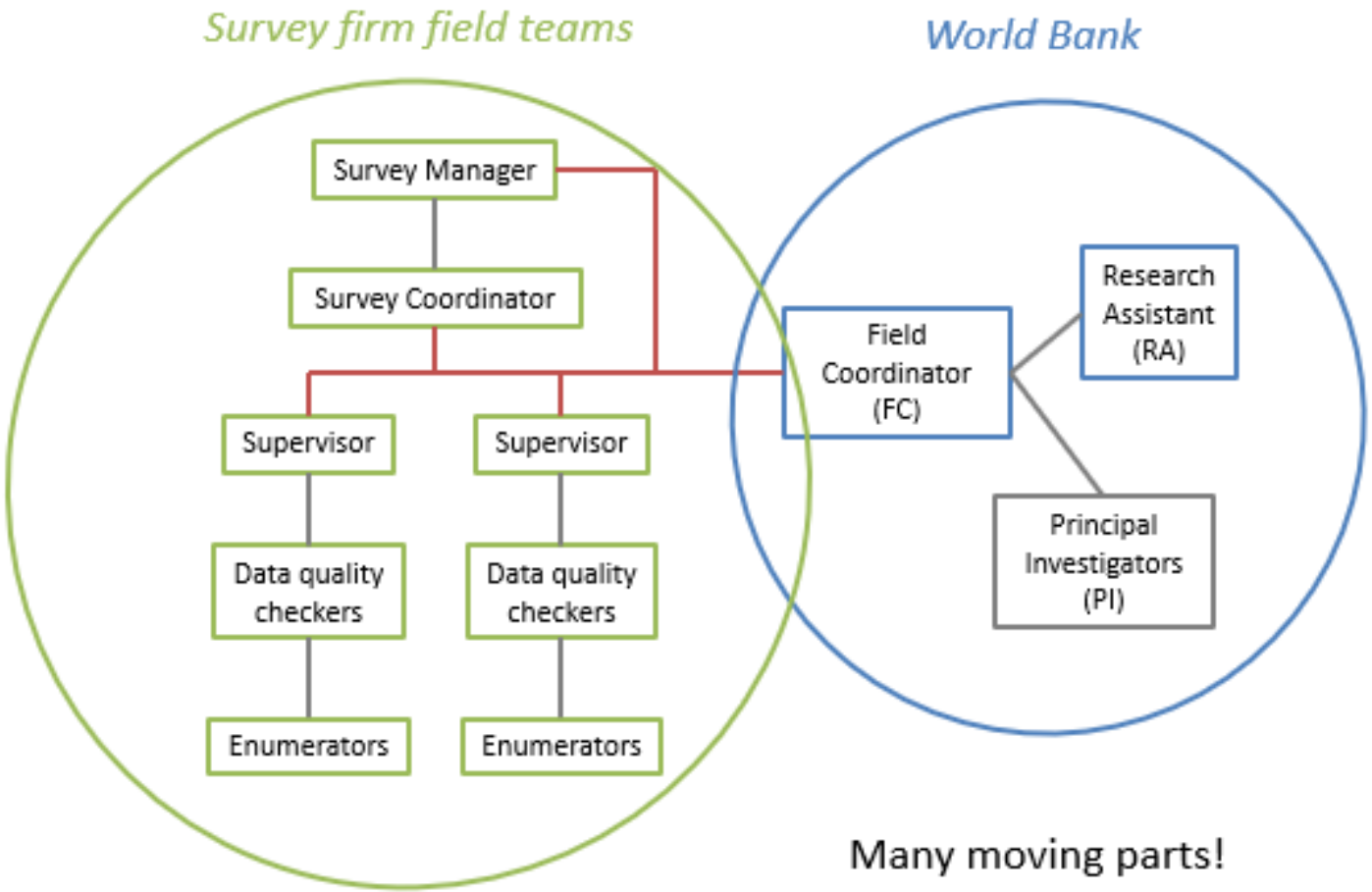
Pre-field	During the field	Post-field
• Survey programming • Enumerator training	• Communication and reporting system • Field monitoring • Minimizing attrition • Real-time data quality checks • Back-checks	• Final field report • Data cleaning

Today's focus: During the Field

Pre-field	During the field	Post-field
• Survey programming • Enumerator training	• Communication and reporting system • Field monitoring • Minimizing attrition • Real-time data quality checks • Back-checks	• Final field report • Data cleaning

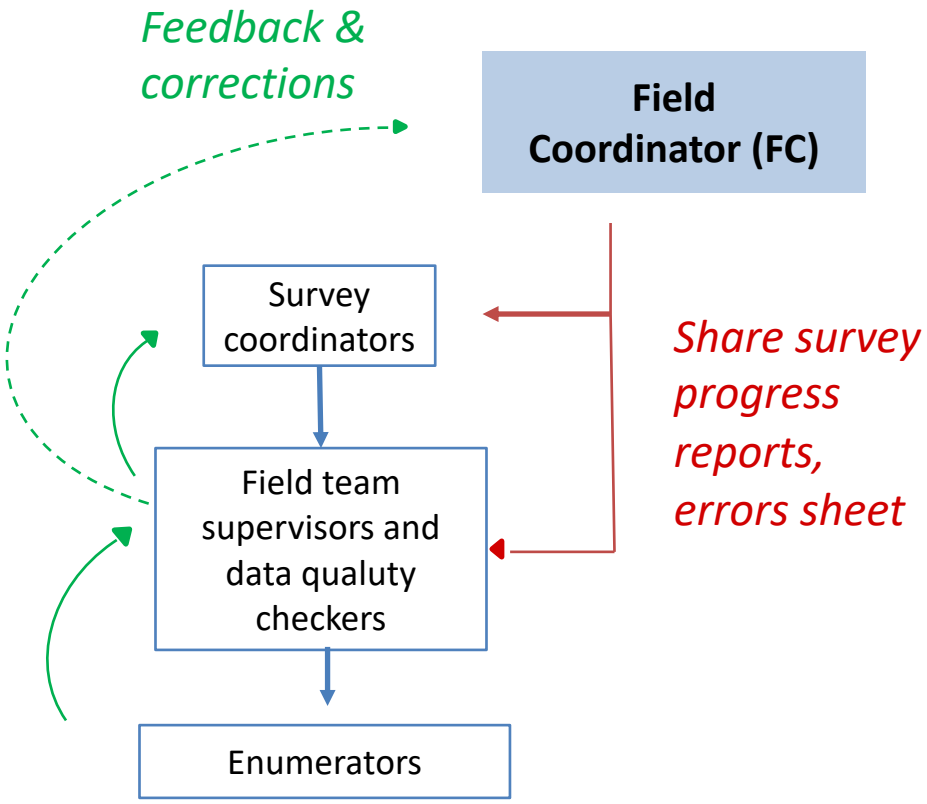
Communication and reporting system

Survey Team Structure



Many moving parts!

Good communication is key!
→ Within the team



- Share reports, observations, error sheets, etc. with the survey coordinator, and/or directly to the field team supervisors (make sure the survey firm agrees with this communication channel)
- Enumerators usually should not report to you, field team supervisors can

Good communication is key!
→ Within the team

- Set up a good system for reporting feedback!
  - Create WhatsApp group with you, survey managers/coordinators and supervisors
    - WhatsApp conversation can be exported & stored on DB
  - Each team meets at the end of the days for feedbacks, sharing experiences on challenges and success
  - Get in touch with your survey manager / coordinators everyday for feedback if you couldn't meet them during the day
  - Keep a diary or log of conversation with teams

Good communication is key!
→ Within the team

- Set up a good system for reporting feedback!
  - Have a shared Dropbox folder with you, the RA, PIs, survey managers/coordinators, and supervisors where
    - HFC outputs are stored
    - Back check files are created
  - Easy to share these files with the team as files auto-update throughout the data collection period
    - No need to send an email with the updated results on a daily basis

Good communication is key!
→ Within the team

- Set up a good system for reporting feedback!
  - Have a shared Dropbox paper with you, the RA, and survey managers/coordinators where
    - The timeline of the project is located and updated in case of changes
    - All tasks related to the data collection are listed
      - Usually assigned to members of the team, and
      - With deadlines to ensure follow-up
    - This can be shared with the PIs as well

Good communication is key!
→ Within the team

Surveys in remote areas can present connectivity issues:

- Cannot send survey forms every day
- Receiving and sending HFCs and backcheck feedback
- Charging tablets
- Phone network

What can you do?

- Train the supervisors well! – many problems don't need FC interaction to resolve
- Set maximum period before having to check in again / sync tablets (not > 48 hours)
- Be creative

An example

- We have an ongoing project in Rwanda where a survey firm has been employed to collect data
- These are the modes of communication set up
  - Shared Dropbox folder where
    - Survey progress is tracked
    - HFC outputs are stored
    - Back check files are created
  - WhatsApp group with the FCs, RA, and survey coordinators
  - Dropbox paper with FCs, RA, and survey coordinators
    - To set up timelines
    - To create to-do lists of tasks and assign to the person in-charge

Dropbox paper example

Tasks before pilot

- Main instrument
  - Modules
    - Up to hiring module
      - [x] ~~@Roshni K to test~~ — Last week RK
    - Plot module
      - [x] ~~@Roshni K to update & test~~ — Last week RK
      - [x] ~~@DIME A to test~~ — Last week DA
      - [x] ~~@DIME A to update translations~~ — Last week DA
    - Transactions module
      - [x] ~~@Roshni K to update & test~~ — Last week RK
      - [x] ~~@DIME A to test~~ — Last week DA
      - [x] ~~@DIME A to update translations~~ — Last week DA
    - Intervention module
      - [x] ~~@Roshni K to update & test~~ — Last week RK
      - [ ] @Christophe N to test — Last week CN
      - [x] ~~@Christophe N to update translations~~ — CN
    - Seasons module
      - [x] ~~@Roshni K to update and test~~ — Last week RK
    - Extension module

High-frequency checks (HFCs)

What are High-frequency checks?

Definition: Checks run on a **daily basis** for **ALL surveys** to provide information about various aspects of the quality of data collected.

- Elements of these checks include
  - The quality of the data
  - Enumerator performance
  - The survey programming (are there programming errors?)
  - The survey progress
- *ipacheck* = IPA Stata package for running high-frequency checks on research data

Categories of High Frequency Checks

Response Quality	Monitor the consistency of responses across the survey instrument and the range within the responses fall. *e.g. standardizing units, outliers*	Revisit the question/ retrain enumerators
Programming	Understand if the questionnaire has been designed and programmed properly. *e.g. sensitive questions, categorizations, calculations*	Survey form updates
Enumerators	Determine if any individual enumerator's data is significantly different from other enumerators' data. *e.g. percentage of "don't know", average interview duration*	Enumerators may need more training
Duplicates and Survey Log	Confirm that all the data from the field is on the survey in a sound manner. *e.g. all data is on the server, duplicates*	Request corrections from field team

High Frequency Checks: Error Source

Is a discrepancy is because:

1. the enumerator recorded the response incorrectly accidently
2. the survey programming has inconsistencies/breaks
3. the enumerator recorded the response incorrectly maliciously (to save time by skipping sections, repeat counts, etc.)
4. the survey form got submitted multiple times to the server
5. 2 enumerators surveyed the same respondent

High Frequency Checks: Dealing with Errors

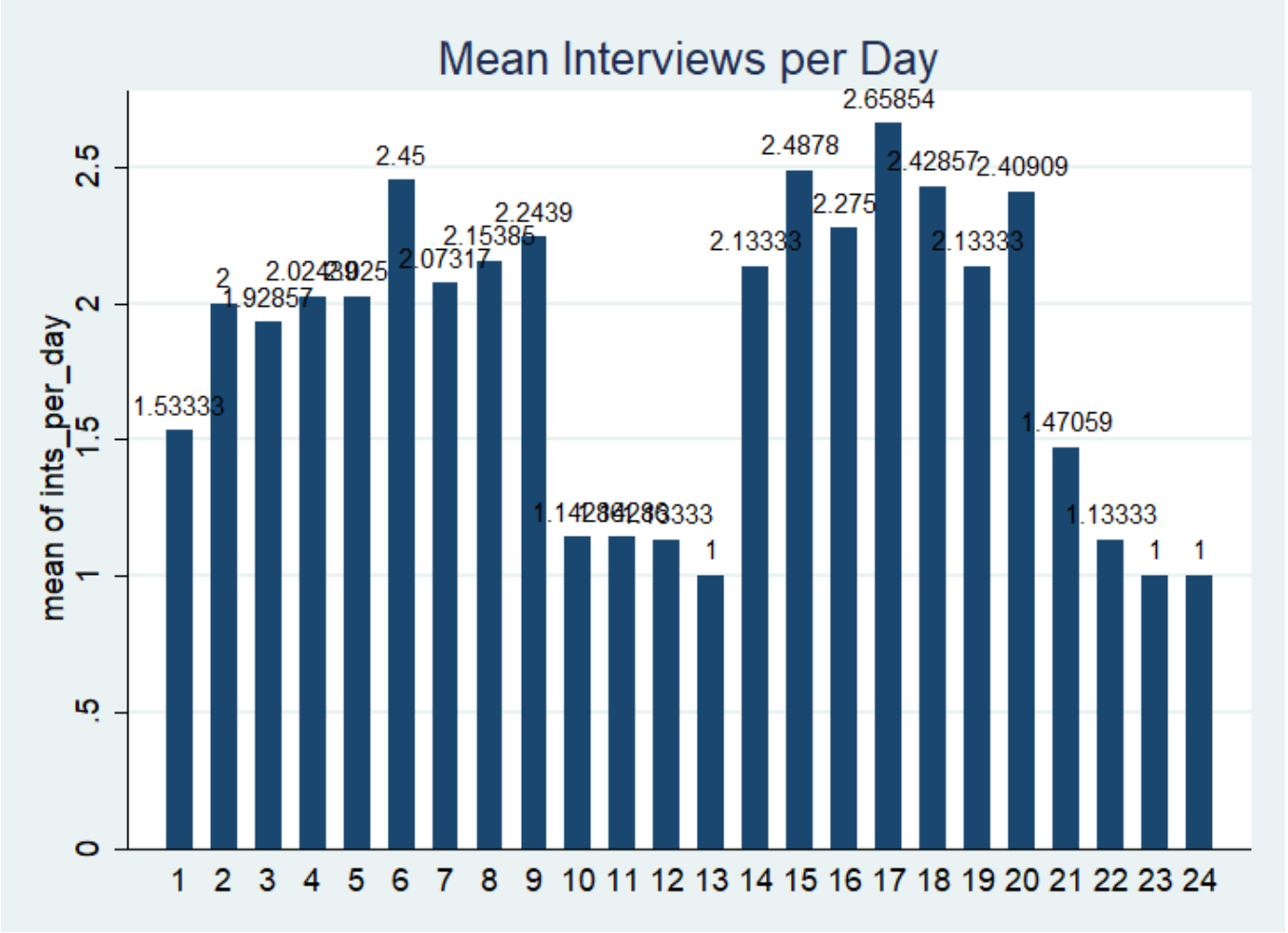
Response Quality	Share daily log of what can be fixed in case of accidental entry errors. Closely monitor in case it's particular enumerators with these errors.
Programming	Update the programming of the survey and test extensively. Ensure the updated version of survey is used on all tablets.
Enumerators	Should start conversations in the management team. Possible actions: re-training, meetings to review protocols and editing the questionnaire
Duplicates and Survey Log	Share a report with the field team and identify reasons for low completion rate. In case of duplicate IDs, identify the correct response to keep.

High Frequency Checks Example

Response quality checks

D_05: HH	ID_03: Enu	tot_flags	tot_all_fla	Flag 1: Ma	Flag 3: HH	Flag 7: mo	Flag 7b: >2	flag_8b	Flag 9 : All
4058	TIBAYIJUK	1	6	0	0	0	0	0	0
2349	MUKANTA	1	4	0	0	0	0	0	0
2123	HAKIZIMA	3	3	0	1	1	1	0	0
2992	SUBIKA Pe	3	3	0	0	1	1	0	0
2102	MAHE Olg	3	3	1	1	0	0	0	0
2083	ISHIMWE I	3	6	0	0	0	0	0	0
4368	HAKIZIMA	3	3	0	0	1	1	0	1
2259	HAKUZIMA	3	4	0	0	1	1	0	0
2055	MAJYAMB	3	3	0	0	1	1	0	1
2853	MUJAWIM	3	3	0	0	0	0	1	1

High Frequency Checks Example



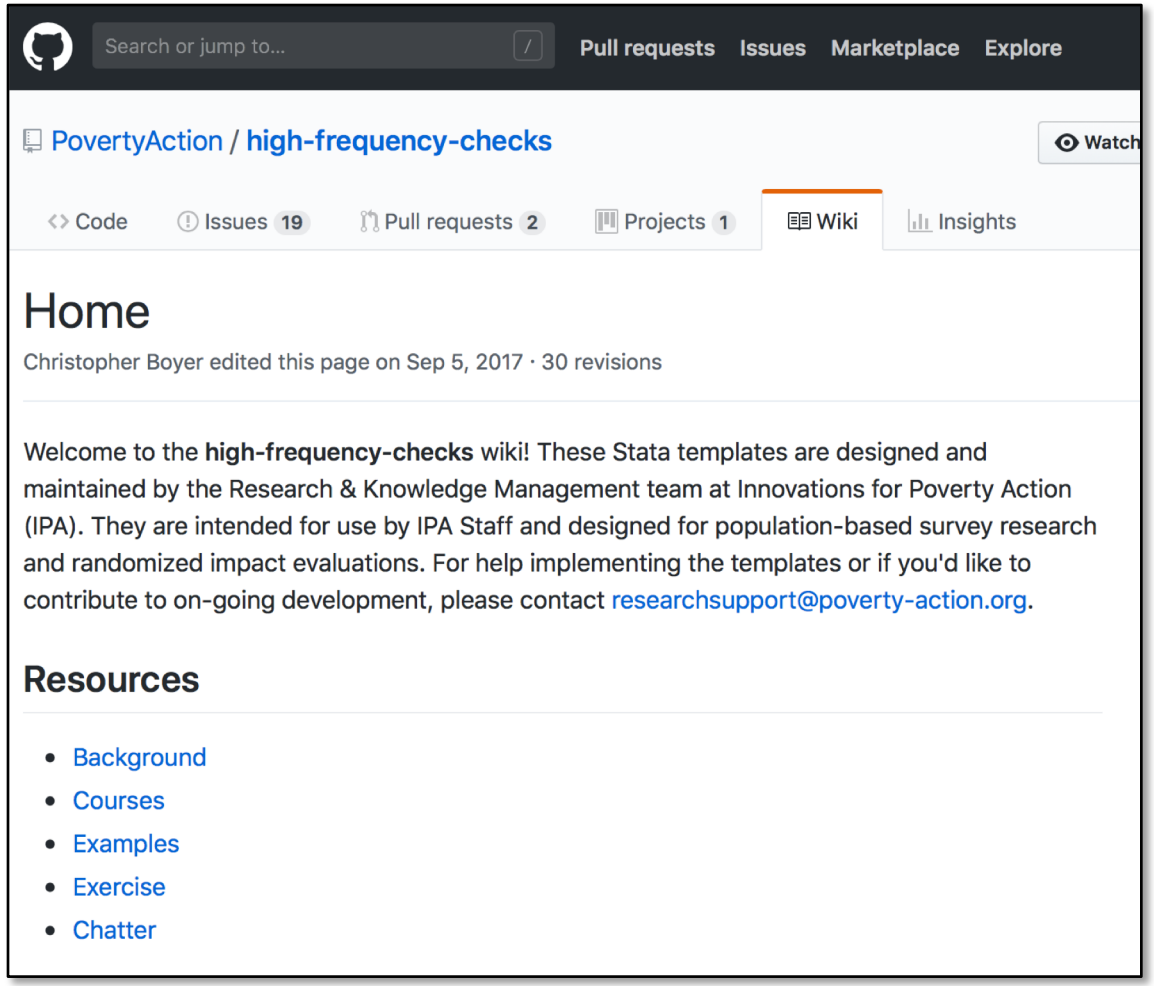
High Frequency Checks Example

Programming checks

id_05	enumerat	Ignore	tech_flag_count	tech_flag_3
2855	23	No	1	1
4174	37	No	1	1
2614	27	No	1	1
2666	42	No	1	1
4480	29	No	1	1
4069	43	No	1	1
2302	46	No	1	1
2639	6	No	1	1
2661	41	No	1	1
2671	28	No	1	1
2335	12	No	1	1

In the office: High-frequency checks

- Done by you:
  - Prepare code / instructions for HFCs before the survey goes to field
  - Monitor flags reporting and address in the field
- Done by the research assistant:
  - Maintain HFC code
  - Daily data download
  - Daily flags reporting
  - This should be a one-click process



https://github.com/PovertyAction/high-frequency-checks

Backchecks

Backchecks

Definition: An experienced enumerator re-visits shortly after the interview and asks selected questions again (10-15 min)

- The answers are compared with the original survey
- Every team and every surveyor must be backchecked as soon as possible, and regularly. Frontload the backcheck sample
- **10-20% of sample, with 20% being administered in the first 2 weeks of field work**
  - Random sub-sample – select it in advance if data is available
  - Include **missing respondents** to verify that your team is not biasing your sample by not tracking hard to find respondents.
  - Observations flagged in other quality tests
  - Surveys of enumerators suspected of cheating

Backcheck Survey

Identifying Respondent and Interview Information

- Identifying information - make sure it's the right person
- Were you interviewed? On which day, what time of day?
- Was the enumerator friendly, annoying, rude, or (s)he was just ok?
- Who was present during the back-check interview
- If the back-check cannot be done, reasons why

Categories of Backcheck Variables

Type 1	Straightforward questions where we expect very little variation. *e.g. age, education, relation to household head, floor type*	Serious enumerator problem
Type 2	Questions where we expect capable enumerators to get the true answer. *e.g. sensitive questions, categorizations, calculations*	Enumerators may need more training
Type 3	Are we asking the question in the right way? Are respondents changing their answers because they don't know? *e.g. It depends on your instrument. They are questions **about the survey**, not surveyor, performance*	Survey needs improvement

Back Checks: Error Source

Is a discrepancy is because:

1. the respondent understood the question both times but decided to change their answer
2. the respondent hadn't understood the question initially
3. the respondent didn't understand the question during the back-check
4. the enumerator recorded the response incorrectly accidently
5. the enumerator recorded the response incorrectly maliciously (to save time by skipping sections, repeat counts, etc.)

→ **Third visit or call in case of high rates of discrepancy.**

Back Checks: Dealing with Errors

Analysis framework

- Set up at the beginning to track deviations early, e.g. “enumerator X always gets his question B wrong” or “question A constantly change across the entire team”
- Decide on error thresholds (is a discrepancy of 5% acceptable? 10%?) and what you corrective measures you will take
- As field work matures, these decisions can be modified if necessary. The goal is to: (1) record discrepancies, (2) maintain a log of errors and (3) set realistic expectations with clear actionable steps
- *bcstats* = J-PAL/IPA Stata package that automatically compares back-check & main survey data and provides an outline for analysis

Back Checks: Dealing with Errors

Type 1	**Overall error rate >10%** →  that there may be systemic problems in the questionnaire or administration. Calculate error rates by team, surveyor and question.
Type 2	**Error rates > 10%** should start conversations in the management team. Possible actions: re-training, meetings to review protocols and editing the questionnaire. Calculate error rates by team, enumerator and question.
Type 3	The error rates by question should be examined. High error rates should be discussed with PIs. Not good practice to change question wording or structure halfway through the survey!

Back Checks Example

Type 1 Q should never ever change, regardless of interviewer, location or time of day.

Examples of these questions include:

- *age at last birthday*
- *whether currently in school*
- *highest level of education*
- *relationship to household head*

Back Checks Example

Team error rates

hhidteam	error_rate	differences	total
1	0.1136	85	748
4	0.0682	57	836
3	0.0527	40	759
2	0.0391	34	869

Back Checks Example

Enumerators with error rates above 10%

enum_id	error_rate	differences	total
314	0.2121	7	33
324	0.1818	8	44
224	0.1688	13	77
122	0.1636	9	55
113	0.1591	14	88
121	0.1558	12	77
223	0.1558	12	77
115	0.1515	15	99
425	0.1414	14	99
123	0.1182	13	110

Back Checks Example

Questions with error rates above 10%

variable	error_rate	differences	total
mod1rel_h_head	0.1336	39	292
mod1highest_school_level	0.1062	31	292

mod1rel_h_head - *How is the head of this household related to you?*

Alternatives to in-field back checks

	Positives	Negatives
Audio recording/audits of survey	• No extra field time • Respondent doesn't need to be interviewed twice • Can randomize the section recorded • Provides information on enumerator behavior	• Someone has to listen to the audio and analyze the data • Audio quality is not guaranteed • Requires additional in informed consent
Back checks via phone	• Cheaper alternative • Quicker to implement	• Low response rate is a concern • May be hard to probe effectively or ask complex questions • Sample can be biased as it depends on which respondents have phone

A quick summary

Summary: Data Quality Checks

High frequency checks (HFCs)	vs.	Backchecks

• 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
		• 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

Summary: Data Quality Check Workflow

- Essential to regularly run HFCs and back checks in first couple of weeks to preempt errors
- 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

Bottom line

- Things are going to go wrong!
- Plan as much as you can
- Establish good communication with your team (internally, at the WB, as well as with implementing partners)
- Keep calm and . . . resolve issues as quickly and as effectively as possible

Thank you!