LEAP Primary Data Collection Workshop

Erick Baumgartner Vrinda Kapoor

July 1st, 2021

Data quality assurance

Collecting data is challenging

IPA
@poverty_action

Here's one example of how dedicated our staff are. Trying to do a follow-up of a @cblatts study in Uganda (*9* years after baseline, 3 yrs since last survey) 1/n

Traduci il Tweet

> **IPA** @poverty_action · 12 dic 2017
> At Innovations for Poverty Action, our dedicated local staff are at the heart of what we do. Help us go the extra mile to find the next effective solution: poverty-action.org/sites/default/...

Collecting data is challenging

- *"Here's one example of how dedicated our staff are. Trying to do a follow-up of a Chris Blattman's (@cblatts) study in Uganda (*9* years after baseline, 3 years since last survey). One enumerator hasn't had power in a while, his phone's out of battery, but he knows who he's looking for."*
- *"He asks around, the guy from the original study sometimes comes by the market in town. Asks around the market. Gets pointed to the guys playing dice over there at the dice table. Goes over, gets friendly with the guys at the dice table. Sure, we know him. He comes by sometimes in the afternoon, maybe around 2PM?"*

Collecting data is challenging

- *"Enumerator settles in across the street where he can keep an eye on the table, sure enough, the guy comes in around 2 to play some dice. Enumerator introduces himself, gets the guy to sit down for a long survey about his income, employment, and everything else Chris Blattman (@cblatts) wants to know about how he's doing 9 years after he got a cash transfer to use for job training."*
- *"Convinces the guy to spend some time (could go into the hours) answering questions. Because losing people to follow up is the enemy of knowledge. Study was in 2008 and these guys tracked down *and* persuaded *84%* of them to participate in 2017. They'll run out every lead, and sometimes spy on a dice table for a day to get every possible respondent."*

Collecting data is challenging

- The story is indeed an example of hard work and persistence, showing how a well done tracking exercise can be implemented. Nevertheless, it is also a sign of the amount of effort demanded from interviewers to find the correct respondent of a survey.
- Interviewing the right individuals is a crucial part of the data collection, and preparing to overcome the possible barriers to this goal should be part of the Data Quality Assurance Plan.

Collecting data is challenging

- Incorrect responses or identification of respondents could be:
  - A mistake...
    - made by the enumerator
    - made by the respondent
  - A fraud
    - made by the enumerator
    - made by the respondent

Collecting data is challenging

- A mistake made by the enumerator
  - Checking identities will normally depend on the information at hand to confirm that the person found is the one that was being tracked. Sometimes enumerators fail to confirm that all of the identity variables were checked, and go on if only some of them were correct. It can even be the case that the right person is found, but some of their info is actually wrong in the tracking data (such as a wrong birthdate or address). This makes it hard for enumerators to decide conclusively on the respondent's identity, and after so much effort, they may be inclined to confirm it.

Collecting data is challenging

- A mistake made by the respondent
  - Sometimes even the respondent may think that he's the person being searched. An homonymous neighbor, someone with the same nickname or other matching characteristics may incline many people (such as the ones playing dice in J-PAL's story) to point the enumerator to a person that's similar, but not exactly the one in the sample.

Collecting data is challenging

- A fraud made by the enumerator
  - There are different degrees of fraud that can be committed by an enumerator, from a plain frauded interview, filled alone in a bar, to an actual interview made with someone with very similar identity characteristics, so that even if the actual respondent was not found, it could be seen as an honest mistake (imagine that, after waiting for the whole day until the respondent would come to play dices in J-PAL's story, the enumerator discovers that the person he's been waiting is not John Wayne from neighborhood A, but rather John Doe from that same neighborhood).

Collecting data is challenging

- A fraud made by the respondent
  - Respondents may also fraud their identities. In data collections where there is a gift for respondents to compensate for the time they have invested in answering the survey, people may try to convince enumerators that they are actually part of the sample. In surveys with adolescents or children, it is also not unusual for them to try to provide false identities to enumerators.

Assuring data quality

- To monitor data quality, it is necessary to have a thorough understanding of the whole data collection process, anticipating which problems may arise
  - This holistic approach is the basis of the "Data Quality Assurance Plan"
  - A general methodological approach to ensure data quality in a data collection

Assuring data quality

- A deep understanding of the data collection and its idiosyncrasies will help spotting possible issues. To gather information, one should...
  - Pilot the survey extensively, documenting eventual issues
  - Understand the sample and the difficulties that may arise to interview or track them
  - Think about the consequences of the data collection choices (interview duration, survey mode, interviewer characteristics, questions asked and appropriateness to the interview location, how to structure spot checks and back-checks, etc.)
  - Think about which check modes are better suited for each of the issues flagged

Assuring data quality

- Based on the anticipated issues, one should decide how each of them will be checked
  - Questionnaire design
  - Interviewing and tracking procedures
  - Spot checks
  - Back-checks
  - High Frequency Checks

Questionnaire design

- Fixable errors should be spotted and fixed during the pilots
- "Constraints" and "relevance" options when programming surveys can help to prevent inconsistencies
  - This process must be well thought, since it can also restrain real responses from outliers (imagine questions such as "How many hours do you spend on the internet on average?" or "How much do you expect to earn after graduation?")
- Previous information (administrative data, previous surveys, etc.) may be used to preload information from respondents, preventing typos, contributing to identity checks and allowing complex questions
- Data from the actual survey can also be preloaded to the Back-check for individual-specific checks

Questionnaire design

- Remember the example regarding revenues in (De Mel et al., 2009)
  - Constraints on inconsistencies could prevent the low correlation between Profits and (Revenues - Costs), but would it correct data quality issues?
- Forcing consistency is not always feasible
  - Recall bias
  - Sensitive questions
  - Different respondents (in a firm, for example)

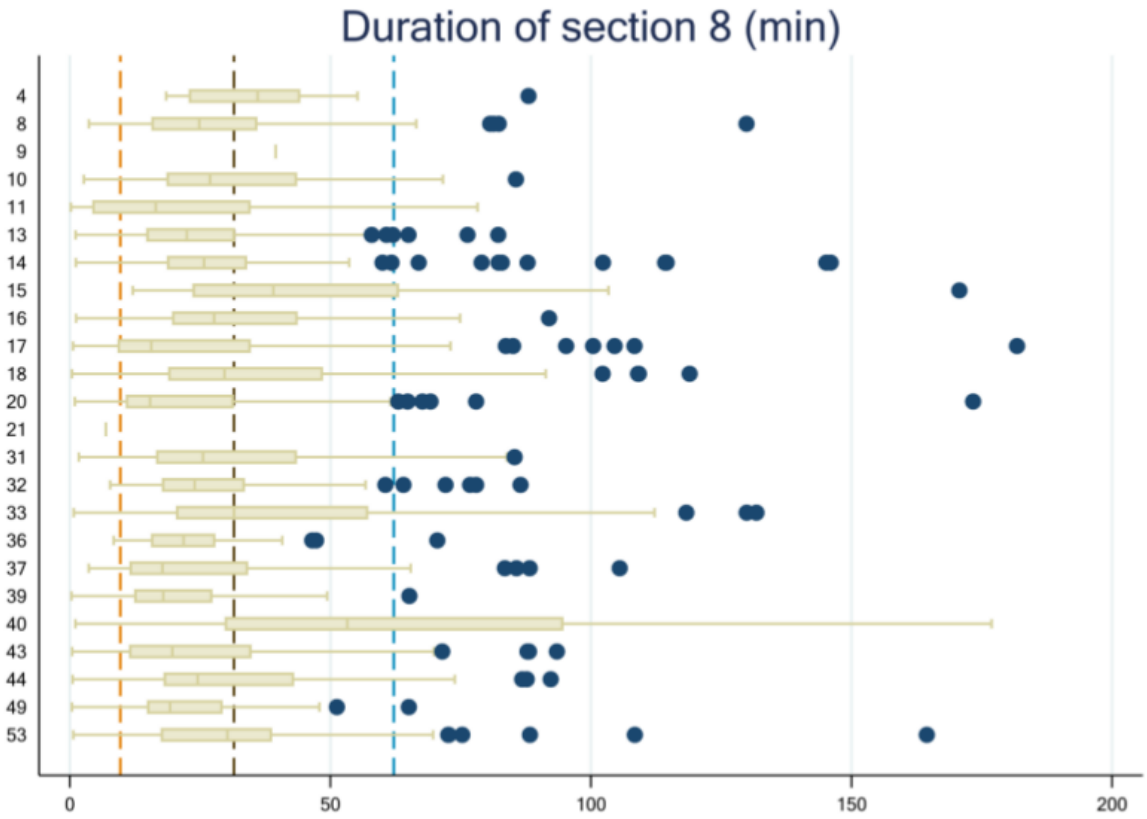
Interviewing and tracking procedure

- Understand the interview and tracking process. What can be checked?
- How do we identify individuals? Is there room for issues?
- Where are interviews being made? Can this cause any biases?
- How difficult is the tracking procedure? Will this create incentives to frauds?
- Even the way in which we approach participants can cause problems

High-frequency checks

- Checks performed continuously
- Allows for quick problem solving, while interviews can still be remembered
- Prompt response to inconsistencies
- Synergy with tracking and logistics

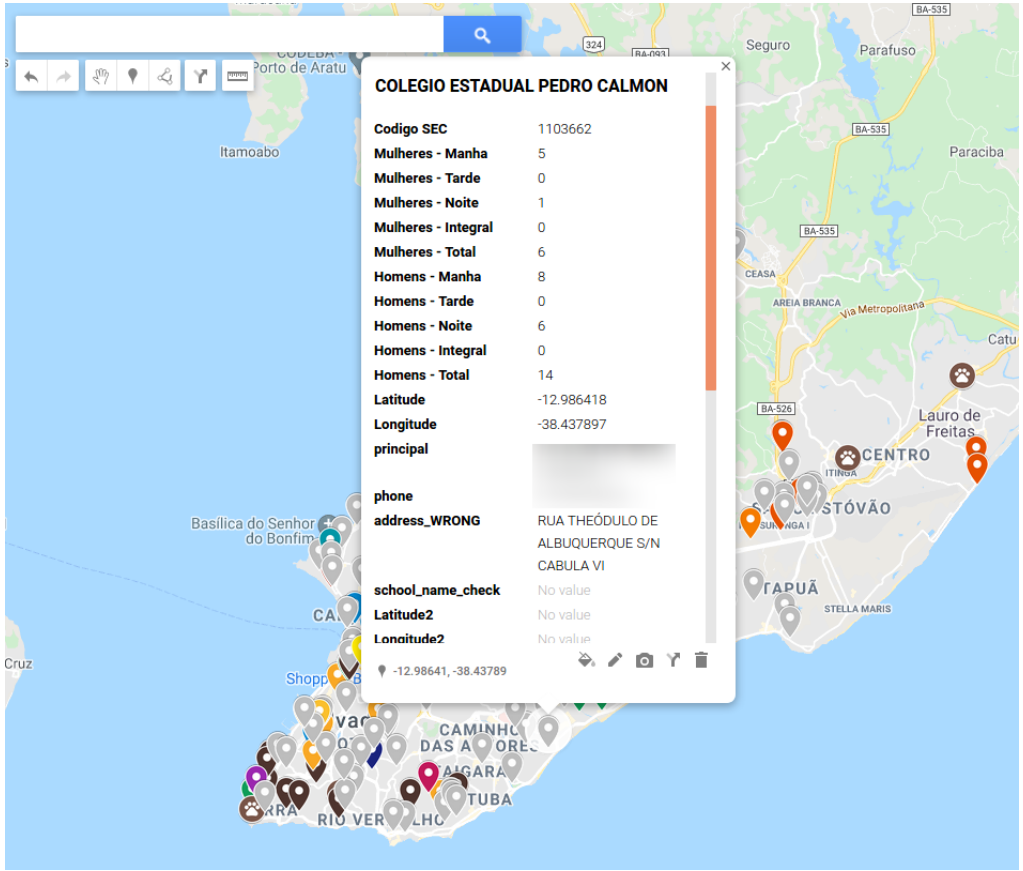
High-frequency checks - examples



High-frequency checks - examples

Interviewer	Team leader	Friends mentioned	Friends mentioned (Q2)	Friends mentioned (Q3)	Number of rounds (Risk preferences game)	Number of rounds (Time preferences game)	Share of respondents with a sexual experience	Average interview duration	Number of interviews
Alexandra	Jeane Barreto	2,77	1,49	0,98	2,03	8,38	0,59	44,10	90
Ana Carolina Falcao	Jeane Barreto	2,89	1,38	0,98	1,89	7,29	0,64	45,89	87
Fabio dos Anjos Bastos	Jeane Barreto	3,11	1,80	0,84	2,95	8,67	0,73	47,55	45
Flavia Couto	Jeane Barreto	3,38	1,61	1,51	2,28	7,27	0,70	50,63	82
Gilcimar Souza de Oliveira	Jeane Barreto	3,71	1,66	1,69	1,97	7,72	0,71	44,54	59
Henrique Saldanha Custodio	Jeane Barreto	3,11	2,22	1,28	3,13	8,94	0,75	44,63	64
Josias Pereira Assuncao	Jeane Barreto	3,13	1,94	2,07	2,45	8,11	0,73	44,86	83
Larisse Favilla Rigaud	Jeane Barreto	3,32	1,59	1,32	2,42	8,11	0,59	57,02	91
Lucila Nascimento da Silva Sampaio	Jeane Barreto	3,29	1,62	1,51	2,17	8,20	0,61	52,90	92
Mirtes Calheiros dos santos	Jeane Barreto	3,00	2,06	2,58	2,13	7,24	0,66	51,43	50
Sample average	-	3,22	1,65	1,54	2,54	8,14	0,69	47,34	

High-frequency checks / Logistics - examples



High-frequency checks / Logistics - examples

	A	B	C	D	E	F	G	H	I	J
1		Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9
2		05/02/2019	06/02/2019	07/02/2019	08/02/2019	11/02/2019	12/02/2019	13/02/2019	14/02/2019	15/02/2019
3	Equipo 01	09DST0069F CESAR USCANGA USCANGA	09DES0159F ELPIDIO LÓPEZ LÓPEZ	09DES0100G LUIS DE CAMOENS	09DES0005C LAURO AGUIRRE	09DES0218E SECUNDARIA GENERAL 218 REPÚBLICA DE ITALIA	09DES0277U LUIS GONZÁLEZ Y GONZÁLEZ	09DST0007T JOSÉ GUADALUPE POSADA AGUILAR	09DES0234W JOSE MANCISIDOR	09DES0123R REPUBLICA ARGENTINA
4	Equipo 02	09DES0017H CONSTITUCIO N DE 1857	09DST0026H CARLOS PELLICER CAMARA	09DES0030B DON BENITO JUÁREZ	09DES0294K JOSÉ PAGES LLERGO	09DES0168N MAXIMILIANO RUIZ CASTAÑEDA	09DES0025Q FERNANDO MONTES DE OCA	09DES0055K REPÚBLICA DE EL SALVADOR	09DES0217F CARLOS CASAS CAMPILLO	LIBRE
5	Equipo 03	09DES0330Z ALTEPECALLI	09DST0034Q LUIS V. MASSIEU	09DES0052N ANTONIO CASO	09DST0112D ESCUELA SECUNDARIA TECNICA 112- REVISITA	09DES0334V SECUNDARIA FEDERAL NO. 334	09DST0079M ESCUELA SECUNDARIA TECNICA 79	09DST0101Y ESCUELA SECUNDARIA TECNICA 101	09DES0321R ACAMPICHTLI	09DST0063L MELCHOR OCAMPO
6	Equipo 04	09DST0065J LUIS LÓPEZ ANTÚNEZ	09DST0104V ING. MARTÍN LÓPEZ RITO	LIBRE	09DES0131Z BELISARIO DOMINGUEZ	09DES0326M SECUNDARIA GENERAL NO. 326 TIEMPO COMPLETO	09DST0030U ING. ALEJANDRO GUILLOT SCHIAFFINO	09DES0282F GILBERTO BORJA NAVARRETE	09DES0241F EMMA GODOY	09DES0122S BRASIL

High-frequency checks / Logistics - examples

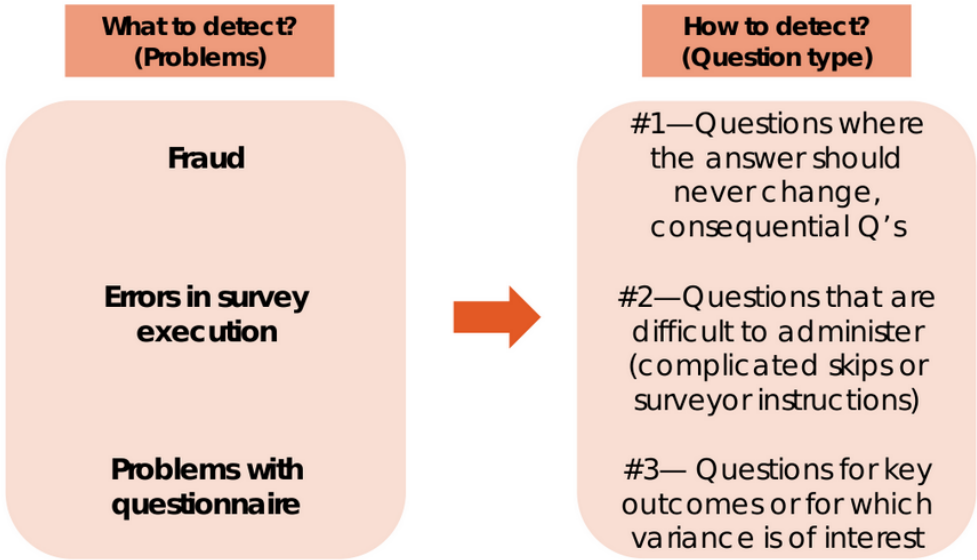
	A	B	C	D	E	F	G	H	I
1	School	Visit	Intervention	1st revisit	2nd revisit	Status	Response rate among Baseline respondents	Response rate among students still enrolled	Students left to be interviewed
2	HEYSER JIM	13/fev/19	6/mar/19	27feb2019		A revisitar	40,68%	37,80%	35
3	CATECPAC	4mar2019	-	03jun2019		A revisitar	40,91%	37,70%	26
4	OVITA A. EL	19/fev/19	12/jun/19	22mar2019	13jun2019	Programar 2a revisita	59,65%	55,56%	23
5	CISCO ZARC	20feb2019	11/jun/19	20mar2019	13jun2019	Programar 2a revisita antes del fin de abril	68,00%	59,70%	16
6	ERVANTES S	28feb2019	-	06jun2019		A revisitar	61,82%	61,76%	21
7	LOS BENITE	18/fev/19	20/mar/19	26feb2019		OK	64,44%	67,92%	16
8	ERNAD PER	6mar2019	-	27feb2019	08may2019	OK	68,33%	60,98%	19
9	HOR OCAMP	12/fev/19	28/fev/19	26feb2019		OK	66,67%	66,67%	15
10	JNDARIA TE	11/fev/19	27/fev/19	21feb2019		OK	70,91%	67,69%	16
11	MOS COMEI	11feb2019	-	02apr2019	09may2019	OK	88,37%	84,31%	5
12	ROVICH PAV	13feb2019	-	06jun2019		A revisitar	80,00%	78,38%	12
13	ASCONCEL	05feb2019	-	21may2019	30may2019	OK	79,10%	74,70%	14
14	IEL A. CHAV	07feb2019	-	25feb2019	06may2019	OK	80,39%	76,19%	10
15	AMPICHTLI (	14/fev/19	11/mar/19	11mar2019	5mar2019	OK	80,95%	78,79%	4
16	UNDARIA TE	08mar2019	-	29mar2019	16may2019	OK	95,00%	94,67%	3
17	ITLAHUAC (	5mar2019	-	28may2019		A revisitar	85,71%	91,30%	5
18	LÓPEZ MAT	7/mar/19	7/jun/19	16may2019	04jun2019	OK	83,33%	81,33%	11
19	BALCÁRCE	27/fev/19	6/mai/19	04apr2019	6may2019	OK	83,93%	84,51%	10

Spot checks

- Visiting interview locations with enumerators is a way to assure that interviewers are following the procedures of the data collection
  - It is particularly important for Field Coordinators to make spot checks in the first phase of the data collection, when enumerators are still adapting to the procedures
  - Provide feedback and solve issues
  - Team leaders can support spot checks throughout the data collection

Back-checks

- A second interview with a short subset of the survey questions allows to check the quality and the legitimacy of the collected data



Back-checks

- Not all inconsistencies represent serious issues
- Think carefully about who will perform it and how it will be performed
- Aim to back-check 10-20% of your sample
- Oversample back-checks in the first stages of data collection, when enumerators might still be improving their knowledge of the procedures
- Include missing respondents
- Include flagged observations

Back-checks

- Questions to...
  - check respondents' identities and interview information
  - detect fraud
  - detect errors in survey execution
  - detect problems with the questionnaire or key outcomes
  - detect problems with burdensome questions, such as multiple loops

Back-checks - example

- Respondent at Baseline: John
  - HH Roster 1: Jack
  - HH Roster 2: Mary
  - HH Roster 3: Susan
  - HH Roster 4: ...

Additional checks

- Audio audits
  - What is the mobile internet service like in the areas being surveyed?
  - Does your team have the capacity to listen to audio audits?
  - Ethical considerations
- Checks using GPS data
- Checks using metadata (SurveyCTO offers options regarding this)
  - Timestamps
  - Light conditions
  - Movement
  - Estimate of the probability of a conversation happening

SurveyCTO - Text audit

- By default, including a text audit field anywhere in your form will record timing information (in seconds) about each field visited while filling out that form. You'll be able to tell when the field first appeared in the form, and the total amount of time that was spent on the field. The text audit .csv file that gets attached to the submission will contain a row for each field.

type	name	appearance
text audit	fieldname	
text audit	fieldname	p=#
text audit	fieldname	eventlog
text audit	fieldname	choices
text audit	fieldname	p=#; eventlog; choices

SurveyCTO - Audio audit

- Audit a random subset of submissions
- Start recording at a specific time
- Start recording at a random time
- Start recording at a specific field
- Stop recording after a specific amount of time
- Stop recording after a specific field

type	name	appearance
audio audit	fieldname	
audio audit	fieldname	p=#
audio audit	fieldname	p=#;s=#;d=#
audio audit	fieldname	p=#;s=#-#;d=#
audio audit	fieldname	p=#;s=startfield;d=endfield

SurveyCTO - Audio audit

- Speed violation: audio-record in response to a certain number of "speed violations" (cases where the enumerator spent less time on fields than specified in the *minimum_seconds* column)

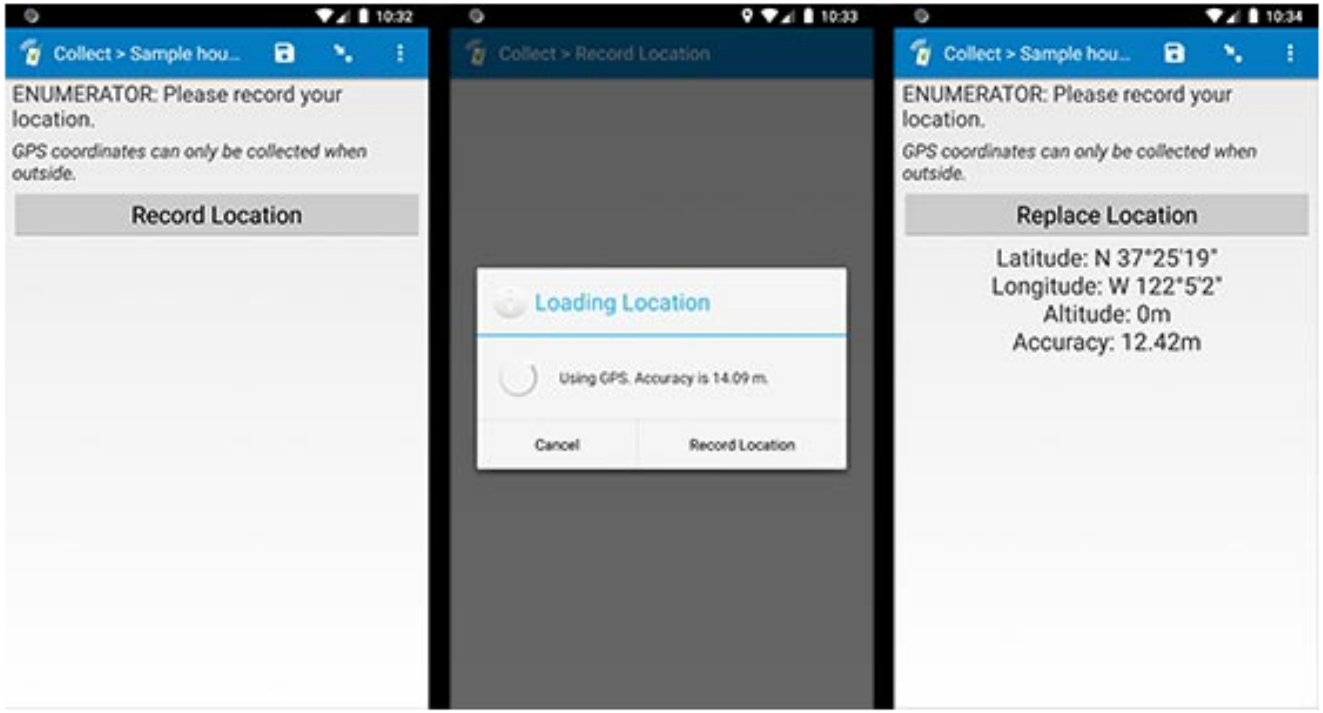
type	name	appearance
speed violations audit	fieldname	v=#; d=#

SurveyCTO - GPS data

- Check it and edit possible mistakes (*maps*, *placement − map*)
- Record it in the background (*background*)
- Select accuracy (*accuracy_threshold*)

type	name	label	appearance	accuracy_threshold
geopoint	fieldname	question text		
geopoint	fieldname	question text	maps	
geopoint	fieldname	question text	placement-map	
geopoint	fieldname	question text	background	
geopoint	fieldname	question text	background	#

SurveyCTO - GPS data



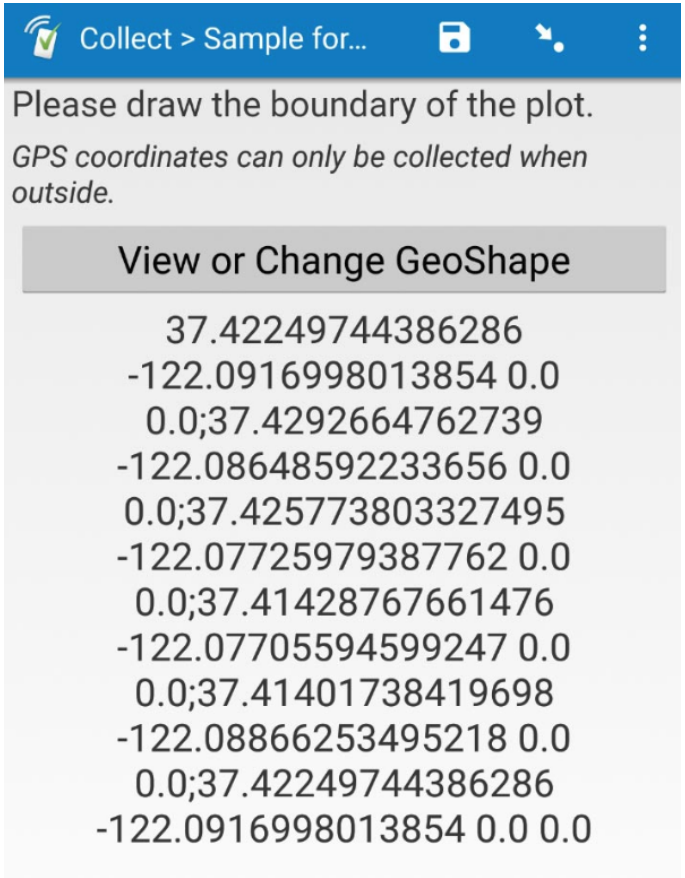
SurveyCTO - GPS data

- Record areas

type	name	label
geoshape	fieldname	question text

SurveyCTO - GPS data

- Record areas



SurveyCTO - GPS data

- Example: Henderson et al. (2020)

Kibera, Nairobi VHR image for 2004 (left) and 2015 (right).

SurveyCTO - GPS data

- Example: Henderson et al. (2020)

Kibera, Nairobi classified buildings for 2004 (left) and 2015 (right) - unchanged (blue), demolished (red), redeveloped (green).

Data quality assurance plan - example

- Follow-up survey
- Sample of High school students interviewed 2 years before
- Around 1,000 respondents had officially dropped out by the time of the Follow-up (2019)
- Survey consisted of in-person interviews
  - School visits
  - Household visits

Data quality assurance plan - example

- High Frequency Checks
- Identity checks based on Baseline data and administrative data sets
  - Names and characteristics of parents and relatives
  - Birthdates
  - Address
  - Enrollment history
- In-person Back-checks made by team leaders
- Phone Back-checks

Data quality assurance plan - example

- Thinking ahead about the possible issues
- Moral hazard
  - Respondents
    - Gifts to respondents
    - Adolescents in schools
  - Enumerators
    - Interviews in dangerous areas
    - Incentives in the tracking phase

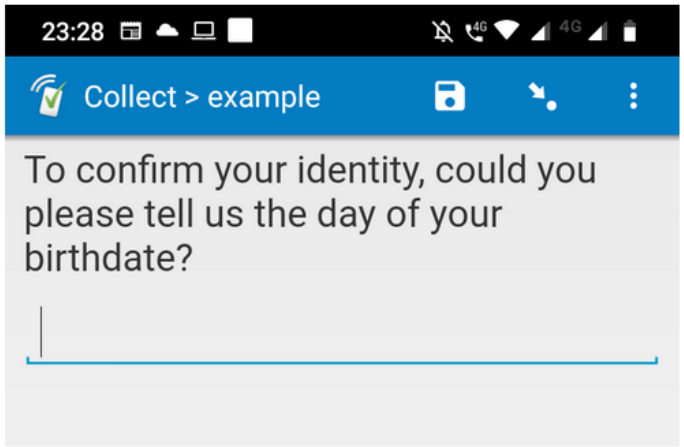
Data quality assurance plan - example

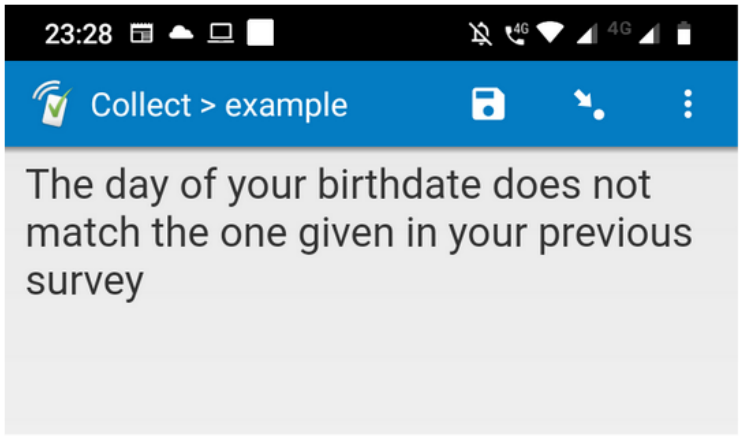
LIST OF REMAINING STUDENTS IN THE SAMPLE

SCHOOL: DONA LEONOR CALMON - TEAM LEADER: Silas

STUDENT	CONTACTS	PHONE	FRIENDS IN 2017	PHONE
NAME: [36] MARIA FRANCISCA [redacted] **(FEM) FORA DA ESCOLA - -**, Turno: -	Fone 2017	71 [redacted]	GLEICIANE MIRANDA CARDOSO	v. 71 [redacted]
END: [redacted]	Antonio Jesus (FATHER)	71 [redacted]	ELLEN SOUZA ALVES	-
DISPON.: -	-	-	RAILANE SANTOS REGIS	-
Dica (2019): -	Tel. Atualizado 1	-	-	-
Escola em 2017: COLEGIO ESTADUAL EDUARDO BAHIANA	Tel. Atualizado 2	-	-	-
OBSERV:				
NAME: [37] MICHELE CRUZ [redacted] **(FEM) FORA DA ESCOLA - -**, Turno: -	Fone 2017	71 [redacted]	LORENA LAIANE MESQUITA DE SOUSA	-
END: [redacted]	Cris Santos (MOTHER)	71 [redacted]	RAQUEL SANTOS SOUZA	at. 71 [redacted]
DISPON.: -	-	-	JEAN VITOR OLIVEIRA SARAIVA	-
Dica (2019): -	Tel. Atualizado 1	-	BRUNA DE JESUS SANTOS	at. 71 [redacted]
Escola em 2017: COLEGIO DA POLICIA MILITAR - CPM JOAO FLORENCIO GOMES	Tel. Atualizado 2	-	ISABEL SANTANA DOS SANTOS	v. 71 [redacted]
OBSERV:				

Data quality assurance plan - example





Exercise - *bcstats* example

```
/*******************************************************************************/
/*******************************************************************************/
/* Example */
webuse          auto, clear; // Loads example data set
gen             enumid = int(_n/10); // Creates enumerator IDs
encode          make, gen(id); // Creates numeric "respondent" ID

tempfile        main;
save            `main'; // Saves main data set

replace         foreign = (1-foreign) if enumid == 5; // Creates mistakes from enumerator 5

tempfile        bc;
save            `bc'; // Saves back-check


/* Sets variables/paths and runs bcstats */
local           output = "${path}bc_diffs2";
local           t1vars = "foreign";
local           t2vars = "headroom";
local           t3vars = "price";
local           ttest = "`t3vars'";
local           id = "id";
local           enum = "enumid";

bcstats,        surveydata(`main') bcdata(`bc') id(`id') enumerator(`enum')
                t1vars(`t1vars') t2vars(`t2vars') t3vars(`t3vars') ttest(`ttest')
                filename(`output') replace;
/*******************************************************************************/
```

Exercise - *bcstats* example

bcstats: Comparing "back checks" in R (a clone of Stata's bcstats)

In vikjam/bcstatsR: Comparing "back checks" in R (a clone of Stata's bcstats)

Description | Usage | Arguments | Details | Value

Description

Comparing "back checks" in R (a clone of Stata's bcstats)

Usage

```
bcstats(surveydata, bcdata, id, enumerator = NA, enumteam = NA,
  backchecker = NA, bcteam = NA, t1vars = NA, t2vars = NA,
  t3vars = NA, ttest = NA, level = 0.95, signrank = NA, lower = FALSE,
  upper = FALSE, nosymbol = FALSE, trim = FALSE, okrange = NA,
  nodiff = NA, exclude = NA)
```

Exercise - *bcstats* example

Displaying variables with high error rates for enumerators with high error rates...

enumid	variable	error_rate	differences	total
5	foreign	1.0000	10	10

Displaying back checks with error rates of at least **30%**...

id	error_rate	differences	total
Audi 5000	1.0000	1	1
Audi Fox	1.0000	1	1
BMW 320i	1.0000	1	1
Datsun 200	1.0000	1	1
Datsun 210	1.0000	1	1
Datsun 510	1.0000	1	1
Datsun 810	1.0000	1	1
Pont. Le Mans	1.0000	1	1
Pont. Phoenix	1.0000	1	1
Pont. Sunbird	1.0000	1	1

Completing **enumerator** checks for type **2** variables...

Displaying variable error rates...

variable	error_rate	differences	total
headroom	0.0000	0	74

Completing **stability** checks for type **3** variables...

Displaying variable error rates...

variable	error_rate	differences	total
price	0.0000	0	74

Exercise - *bcstats* example

ttest for **price**:

Paired t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
price	**74**	**6165.257**	**342.8719**	**2949.496**	**5481.914**	**6848.6**
bc_price	**74**	**6165.257**	**342.8719**	**2949.496**	**5481.914**	**6848.6**
diff	**74**	**0**	**0**	**0**	**0**	**0**

mean(diff) = mean(**price** - **bc_price**) t = .

Ho: mean(diff) = 0 degrees of freedom = **73**

Ha: mean(diff) < 0 Pr(T < t) = .

Ha: mean(diff) != 0 Pr(|T| > |t|) = .

Ha: mean(diff) > 0 Pr(T > t) = .

Exercise - *bcstats* example

id	enumid	type	variable	survey	back_check
Audi 5000	5	type 1	foreign	Foreign	Domestic
Audi Fox	5	type 1	foreign	Foreign	Domestic
BMW 320i	5	type 1	foreign	Foreign	Domestic
Datsun 200	5	type 1	foreign	Foreign	Domestic
Datsun 210	5	type 1	foreign	Foreign	Domestic
Datsun 510	5	type 1	foreign	Foreign	Domestic
Datsun 810	5	type 1	foreign	Foreign	Domestic
Pont. Le Mans	5	type 1	foreign	Domestic	Foreign
Pont. Phoenix	5	type 1	foreign	Domestic	Foreign
Pont. Sunbird	5	type 1	foreign	Domestic	Foreign

Thank you!