LEAP Primary Data Collection Workshop

Erick Baumgartner   Vrinda Kapoor

28 June, 2021

Introduction

- ▶ If you have any questions or comments, send us a message!
  - ▶ erick.baumgartner@unibocconi.it
  - ▶ vrinda.kapoor@unibocconi.it
- ▶ There are also many additional sources on data collection best practices
  - ▶ DIME
    - ▶ Development Research in Practice - The Dime Analytics Handbook
    - ▶ Online course: Manage Succesful Impact Evaluations Surveys
  - ▶ J-PAL
  - ▶ BITSS

Preparing for a successful survey

Survey workflow
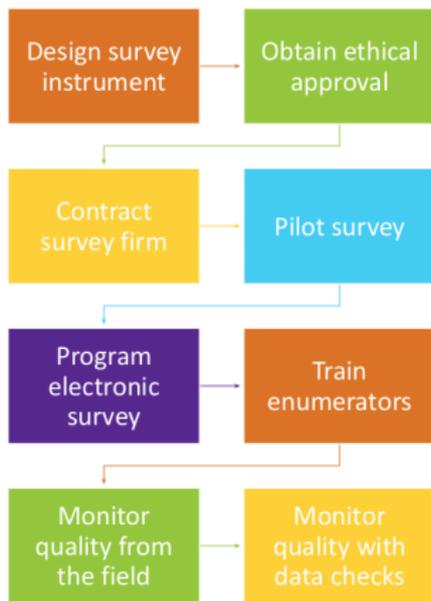
Have a clear vision of the complete process

- Once data collection starts, you'll have to manage a continuous flow of information
  - Every issue will have consequences for the next steps
  - The later you identify problems, the harder it will be to solve them
  - Try to imagine everything that could go wrong

Know your objective

- It is imperative that the survey has a clear objective and that the survey instrument and protocols are designed to achieve that objective
  - Does the questionnaire address the broader research questions?
    - It will be hard to correct any mistake (Restarting the process, IRB, Pre-Analysis Plan)
  - Does the survey design allow us to collect the right data at the right time?
  - Do the survey protocols ensure that the correct respondents will be identified and interviewed?

Know your objective

- "Research with adolescents: issues surrounding consent" (Berk Özler's post on the Development Impact blog)
  - Every aspect of the data collection will have consequences to its results
  - Consent bias, feasibility, etc.
- "What if Christopher Columbus had been funded by the Gates Foundation?"
  - Rigorous Evaluation finds no effect of Going West approach on Reach the Indies

Context matters

- The survey design (from the questionnaire to the implementation logistics) must account for, and reflect, the peculiarities, requirements, customs, and limitations of the local context
  - All stages of survey preparation should be preceded by the question: "will this work in this context?"

Define who does what

- Clear roles, responsibilities, and expected deliverables of all involved, as well as good planning and regular communication between all relevant parties, is crucial
  - Does everyone know what role they play and what is expected of them, and when?
  - Is the communication between survey partners frequent and open enough to allow for honest and constructive engagement?

Define who does what

- ▶ Research assistant
  - ▶ Program questionnaire
  - ▶ Perform data quality checks
  - ▶ Feedbacks on revisits, improvements, issues with interviews
- ▶ Field coordinator
  - ▶ Coordination with partners (survey firm, enumerators, implementing partners, research team, etc)
  - ▶ Training enumerators
  - ▶ Assuring the impact evaluation methodology
  - ▶ Logistics
  - ▶ Many unforeseeable issues

Examples

- Imagine that the business owners in our sample are distributed across the whole country.
    - How could we structure this data collection?
    - In the pilots, the survey had an average duration of around 90 minutes.
    - The survey firm is based in Ouagadougou

Examples

- ▶ How many interviews can be performed in one day?
- ▶ Should enumerators stay until they find every single respondent in our sample?
- ▶ Can enumerators perform back-checks?
- ▶ How long should we wait to send a sweeper?
- ▶ How long should we wait to send a back-checker?

Examples

	Business owners to interview	Enumerator days	Expected cost
Banfora	5	2	US$ 80
Bobo-Dioulasso	7	3	US$ 120
Diébougou	3	1	US$ 40
City 4	3	1	US$ 40
City 5	8	3	US$ 120
City 6	2	1	US$ 40
Ouagadougou	7	3	US$ 120
			US$ 560

Expected costs

Logistics

Examples

	Business owners to interview	Enumerator days	Back-checker days	Expected cost
Banfora	5	2	1	US$ 120
Bobo-Dioulasso	7	3	1	US$ 160
Diébougou	3	1	1	US$ 80
City 4	3	1	1	US$ 80
City 5	8	3	1	US$ 160
City 6	2	1	1	US$ 80
Ouagadougou	7	3	1	US$ 160
				US$ 880

Expected costs

Logistics

Examples

	Business owners to interview	Enumerator days	Back-checker days	Sweeper days	Expected cost
Banfora	5	2	1	1	US$ 160
Bobo-Dioulasso	7	3	1	1	US$ 200
Diébougou	3	1	1	1	US$ 120
City 4	3	1	1	1	US$ 120
City 5	8	3	1	2	US$ 240
City 6	2	1	1	1	US$ 120
Ouagadougou	7	3	1	1	US$ 200
					US$ 1160

Expected costs

Logistics

Examples

Planning visits
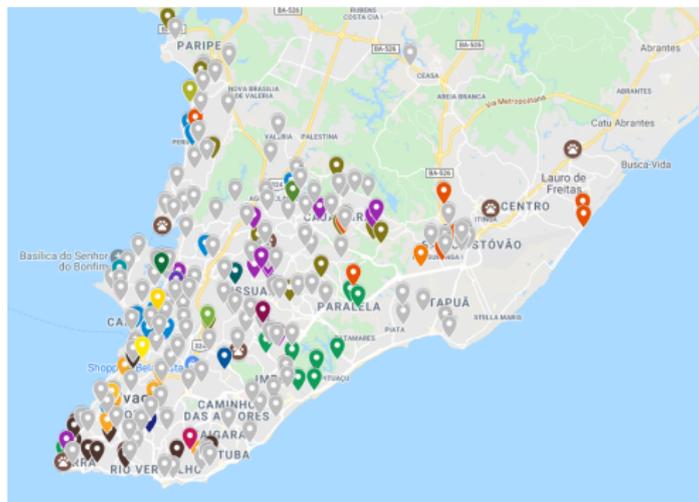
Examples

Planning visits

Examples
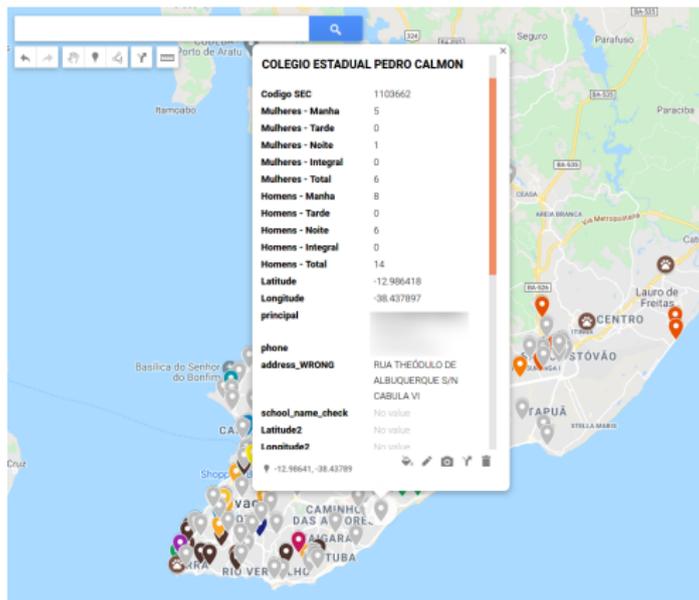
Planning visits

Designing and piloting a questionnaire

Designing a questionnaire

- ▶ It's hard to ask questions correctly and reach high quality standards in data collection
- ▶ Measuring microenterprise profits: Must we ask how the sausage is made? (de Mel, McKenzie and Woodruff 2009)
  - ▶ "Otto von Bismarck famously remarked that people who like sausages and laws should not see how either of them is made. Economists may wish to say the same about profit data from microenterprises in developing countries."

Designing a questionnaire

- ▶ Survey with microenterprises in Sri Lanka between 2005 and 2007
  - ▶ In the baseline surveys firm owners were asked directly for their profits in the last month, and also to report revenue and expenses.
  - ▶ The level of reported profits is substantially higher than the level of reported revenues minus expenses at both the mean and the median.
  - ▶ Moreover, the correlation between reported profits and reported revenue minus expenses is only 0.2–0.3, and 30% of firms have negative revenue minus expenses.

Designing a questionnaire

- "We examine how far one can go towards reconciling the difference between profits and revenue–expenses through asking more detailed survey questions and through better matching of revenues with the expenses incurred to meet these revenues."
- Corrections bring the levels of the two measures much closer to one another, and lead to a marked improvement in the correlations — to 0.61–0.73.

Designing a questionnaire

- Sources of incorrect reports
  - Confusion between business and household expenses
  - Mismatching of revenues and expenses
  - Recall errors
  - Deliberate underreporting

Designing a questionnaire

Table 9
If the truth is 10,000, how much would owners report?

	Revenue			Expenses		Profits	
	KMS-all	KMS-2 day	SLMS	KMS	SLMS	KMS	SLMS
Mean	6744	6668	8204	12,254	11,417	NA	7801
10th percentile	4500	5000	6000	10,000	8000	NA	5000
25th percentile	5000	5500	7000	10,000	10,000	NA	6000
50th percentile	7000	7000	8000	12,000	10,000	NA	8000
75th percentile	8000	7500	10,000	15,000	13,000	NA	10,000
90th percentile	10,000	10,000	10,000	15,000	15,000	NA	10,000

Note: KMS-all indicates all KMS firms; KMS-2 day refers to the 58 KMS firms receiving visits every 2 days.

Assessing underreporting

Designing a questionnaire

- "The majority of firms think that revenues are underreported, with a median level of underreporting of 20% in one sample and 30% in the other."
- "We had research assistants observe firms in the second sample 15 to 16 times during a month and record transactions, and use this to estimate actual revenues for these firms. The reported revenues of firms are 31% lower than we estimate, confirming the level of underreporting suggested in the self-reports."

Piloting a questionnaire

- Pre-pilot
- Content-focused pilot
- Data-focused pilot

Pre-pilot

- Answer broad questions about qualitative factors like questionnaire design and the context for conducting the study.
- Learn more about how people think to see what structure and flow makes sense. For example, do people think about their input use at plot-level? By crop? Overall?

Content-focused pilot

- ▶ Refine the order and wording of specific questions, the overall structure of the questionnaire, and translations.
- ▶ Check that the answer choices are comprehensive, that is, they cover all possibilities.
- ▶ Flag any sensitive questions.
- ▶ Check how long it takes to answer the questions. Also check how answers differ between respondents (response variance)

Data-focused pilot

▶ Check if the programmed instrument displays questions in the correct order, and follows the correct patterns (such as a group of questions which will always appear together, or will repeat).

▶ Load a sample data set (based on pilot interviews). Check if the data set has any missing fields. Perform all data quality checks, like back checks and high frequency checks.

Checklist

	Initials	#No	Checklist Item
1. All stages			
[_]		1.1	Have you identified a sufficient number of qualified interviewers?
[_]		1.2	Have you trained the interviewers on the survey instrument?
[_]		1.3	Have you identified a comparable area and population for the pilot?
[_]		1.4	Have you secured all approvals / letters of support needed in your context?
[_]		1.5	Has someone contacted the local leaders in the pilot area to inform them of planned survey activity?
[_]		1.6	Will the team be staying overnight in the pilot area? If so, do you have necessary permissions to travel (from government, World Bank, etc)?
[_]		1.7	Has someone taken care of the logistics (e.g. car rentals, meals or per diem for interviewers)?
[_]		1.8	Do you have a venue reserved for training the interviewers?
2. Pre-Pilot			
	Initials	#No	Checklist Item
[_]		2.1	Do you have a set protocol for identifying participants in focus group discussions?
[_]		2.2	Do you have a form prepared for interviewers to record qualitative observations and notes from discussion?
[_]		2.3	Have you tested interviewers' note-taking abilities during the training and provided feedback on content and handwriting?
[_]		2.4	If you do not speak the local language(s), do you have a tried-and-true research assistant to accompany you to translate?

Checklist - Pre-pilot

Checklist

	Initials	#No	Checklist Item
3. Content-Focused Pilot			
[_]		3.1	Is the questionnaire you are piloting fully translated into the local language?
[_]		3.2	Is the instrument formatted for printing? Make sure there are sufficient margins on all pages for taking notes.
[_]		3.3	Did you print enough copies of the instrument for all interviewers and all people accompanying interviews to have a copy for each planned interview?
[_]		3.4	Does the survey instrument include both the research team language and local language (where different)? If not, remember to print out copies in each language as applicable.
[_]		3.5	Do you have access to a printer to print revisions in real time if significant changes are needed?
[_]		3.6	Have you instructed interviewers (or observers) to record the start and end time for each module?

Checklist - Content-focused pilot

Checklist

	Initials	#No	Checklist Item
			4. Data-Focused Pilot
[_]		4.1	Have you office-tested the final version of the programmed instrument for any bugs, and to ensure that all questions appear as expected and reflect the final translation?
[_]		4.2	Have you set up a SurveyCTO server for the pilot? (For DIME members: Remember, no data can be uploaded to the DIME Test server.)
[_]		4.2.1	Is the name for the pilot form on the server clearly distinguishable from the final survey?
[_]		4.2.2	Have you assigned a form id unique to the pilot? (to avoid confusing pilot data with final data)
[_]		4.2.3	Do all team members who need it have log-on information?
[_]		4.3	Is the SurveyCTO Collect app updated to the version of SurveyCTO your server is running? (See the 'Collect' tab on your server for instructions.)
[_]		4.3.1	Best to update your server and app to the latest version of SurveyCTO before starting the pilot.
[_]		4.3.2	Once you have piloted, avoid updating the app even if a new version is released, to avoid compatibility issues.
[_]		4.4	Are all tablets / phones running the most updated version of their operating system (OS)? (or at a minimum, are they all running exactly the same OS?)
[_]		4.5	Are all tablets / phones set to the correct date and time?
[_]		4.6	Is the pilot-form downloaded and ready-to-go on all tablets?
[_]		4.7	Are all tablets fully charged? Do you have battery packs with you in case any batteries run out?
[_]		4.8	Do you have a couple of paper copies of the survey, for observers to use, and/or as a last resort in case of unsolvable tablet problems?
[_]		4.9	Have you built in time to the field plan to make any required revisions to the programming and re-download the revised forms on all tablets?
[_]		4.10	Have you set up Stata do-files for importing and labeling data?
[_]		4.11	Have you set up a Stata do-file for running high frequency checks?

Checklist - Data-focused pilot

Piloting a questionnaire

- Data from the pilots can be used as an information source for power calculations
- Field testing: essential for enumerator training

Survey mode

- ▶ The way in which data is collected may also have an effect on responses
    - ▶ Electronic (CAPI)
    - ▶ Telephone (CATI)
        - ▶ Dabalen, Andrew, et al. Mobile phone panel surveys in developing countries: a practical guide for microdata collection. The World Bank, 2016.
        - ▶ Garlick, Rob; Kate Orkin, and Simon Quinn. Call me maybe: experimental evidence on using mobile phone surveys to survey microenterprises. CSAE Working Paper WPS/2016-14 (2015)
    - ▶ Web-based (CAWI)
    - ▶ SMS-based
    - ▶ Video

Survey mode - CAPI x CATI

Pros

- ▶ Simplifies access to respondents
  - ▶ Transportation
  - ▶ Time availability
  - ▶ Security
- ▶ Cost effective
- ▶ Bias
  - ▶ Higher probability of reaching some of the difficult respondents

Survey mode - CAPI x CATI

Cons

- ▶ Difficulties to check interviews
    - ▶ Some options may not be available (Audio, GPS)
    - ▶ Identity checks
    - ▶ Frauds
- ▶ Bias concerns
    - ▶ Higher demand of previous contact information
- ▶ Concerns regarding attention to questions
- ▶ Slower pace (higher interview duration compared to CAPI)
- ▶ Higher probability of refusals

Thank you!