Primary Data Collection

STEG course on Data in Macro Development

Meredith Startz

June 7, 2024

Primary data collection (for macro development)

- What is “primary data”?

Primary data collection (for macro development)

- What is "primary data"?
  - I mean "data you collect yourself"

Primary data collection (for macro development)

- What is "primary data"?
  - I mean "data you collect yourself"
  - vs. . . . any other source

Primary data collection (for macro development)

- What is "primary data"?
  - I mean "data you collect yourself"
  - vs. . . . any other source
- The "ugly duckling" in this course – one of these things is not like the others!

Primary data collection (for macro development)

- What is "primary data"?
  - I mean "data you collect yourself"
  - vs. . . . any other source
- The "ugly duckling" in this course – one of these things is not like the others!
- Often related to running RCTs in development, but I'll focus just on the data collection, not running experiments
- Understanding data collection is useful, even if you don't do it yourself
  - Important overlap with using survey data collected by someone else for research purposes (e.g. LSMS)
  - Makes you a more insightful, careful user of administrative data sources

Plan for this lecture

- A different approach from the other lectures in this course
  - By definition, I can't tell you about commonly used data sources!
- Instead, I will cover:
  1. When and why to do primary data collection (or not)
  2. Two examples in a macro development setting
     1. The Value of Face-to-Face (my JMP, traders in Nigeria)
     2. Achieving Scale Collectively (Bassi et al. 2022, manufacturers in Uganda)
  3. Technical and practical issues
     1. Sampling
     2. Questionnaire design
     3. Preparation and management
     4. Data quality

When might you want to collect data?

1. When something is not covered well by existing data sources
   - A population or part of the economy that isn't covered (e.g. informal, small firms)
   - A topic that isn't covered (could be anything...)
2. To evaluate the causal impact of something through an RCT or evaluation of a new policy
3. To learn about how the world (really) works
   - Often, this only requires qualitative, exploratory interviewing, not full data collection

When might you want to collect data?

1. When something is not covered well by existing data sources
   - A population or part of the economy that isn't covered (e.g. informal, small firms)
   - A topic that isn't covered (could be anything...)
2. To evaluate the causal impact of something through an RCT or evaluation of a new policy
3. To learn about how the world (really) works
   - Often, this only requires qualitative, exploratory interviewing, not full data collection
   - STEG PhD grants: *"Grants will also support travel to field sites, even when secondary data is utilised. We view this kind of travel (with the possibilities for field visits and conversations with policy makers) as particularly important for researchers who lack prior experience in the countries that they intend to study."*

Examples of using primary data collection

- Population not covered
  - F2F – wholesale traders
- Topic not covered
  - F2F – traveling to supplier locations when sourcing goods
  - ASC – rental of specific production equipment
- To get a causal estimate following an RCT – e.g. experimentally varying costs of seasonal migration
  - Bryan, Gharad, Shyamal Chowdhury, and Ahmed Mushfiq Mobarak. "Underinvestment in a profitable technology: The case of seasonal migration in Bangladesh." Econometrica 82.5 (2014): 1671-1748.
  - Lagakos, David, Ahmed Mushfiq Mobarak, and Michael E. Waugh. "The welfare effects of encouraging rural–urban migration." Econometrica 91.3 (2023): 803-837.

Pros of data collection

- You're free!** **in a highly constrained, costly sense...
  - Not limited to studying what others already find interesting/important (and therefore have collected data about)
  - Can be more creative, more likely to hit on something really new
  - This is high risk, high return
- When you use data someone else collected
  - Hard to find data or variation not already used by researchers
  - Data collected for non-research purposes (e.g. government, private sector):
    - ★ Measures only things they are interested in
    - ★ Covers populations that are of interest and/or self-selected
    - ★ Suffers from any potential measurement or reporting biases that source is prone to

Pros of data collection

- You're free!** **in a highly constrained, costly sense...
  - Not limited to studying what others already find interesting/important (and therefore have collected data about)
  - Can be more creative, more likely to hit on something really new
  - This is high risk, high return
- When you use data someone else collected
  - Hard to find data or variation not already used by researchers
  - Data collected for non-research purposes (e.g. government, private sector):
    - ★ Measures only things they are interested in
    - ★ Covers populations that are of interest and/or self-selected
    - ★ Suffers from any potential measurement or reporting biases that source is prone to
- Consumption value
  - You can be more grounded, team-oriented, anthropological in your work, if that is your style

Cons of data collection

- **Time** – Your colleagues' work starts where your data collection ends
- **Risk** – Macro (pandemic, war), micro (logistics, HR, fraud, budget overruns), varying returns (your results might not be interesting, or well-powered) – hard to "fail fast"
- **Financial cost** – Direct, but also downstream consequesces thereof, e.g. small samples
- **Different skills** – Requires a totally different set of professional and technical skills
  - Survey design, data quality and management, human management, logistics and operations in challenging settings
  - Don't underestimate this!!

Non-cons of data collection (that aren't really cons)

- Economics audiences often view "admin" data as objective truth, whereas primary data is...
  - "Small"
  - Not representative
  - "Self-reported", subject to recall, misreporting biases

Non-cons of data collection (that aren't really cons)

- Economics audiences often view "admin" data as objective truth, whereas primary data is...
  - "Small"
  - Not representative
  - "Self-reported", subject to recall, misreporting biases
- But, data comes from somewhere! It's (almost) always self-reported
- "Admin" data can have all these problems – often worse. E.g.:
  - Min firm size for formal registration, or min transaction size for VAT
  - Recall and misreporting biases in tax or trade data
  - App data only reflects people with smart phones
  - Grant (2023) – data classifications reflect government purposes, e.g. codes in trade data are "finer" for higher tariff goods

Non-cons of data collection (that aren't really cons)

- Economics audiences often view "admin" data as objective truth, whereas primary data is...
  - "Small"
  - Not representative
  - "Self-reported", subject to recall, misreporting biases
- But, data comes from somewhere! It's (almost) always self-reported
- "Admin" data can have all these problems – often worse. E.g.:
  - Min firm size for formal registration, or min transaction size for VAT
  - Recall and misreporting biases in tax or trade data
  - App data only reflects people with smart phones
  - Grant (2023) – data classifications reflect government purposes, e.g. codes in trade data are "finer" for higher tariff goods
- Concerns are good if they encourage you to think carefully about your data generating process. Just apply the same standards to data you download!

Upshots for big-picture strategy

- Don't do it unless you're clear about why, and take on the costs and risks with open eyes
  - Maybe you just need survey data, rather than admin
  - Some existing survey data sources are great, e.g. LSMS -- DON'T recreate the wheel!
- Primary data collection is more common in micro than macro development
  - Macro traditionally wants to quantify aggregates, is less concerned with causal identification
  - Increasingly, more concern about causality, and more microfoundations in macro/equilibrium models
  - $\Rightarrow$ Convergence between macro topics and micro development methods and tools
    - ★ Trade, firms, migration, labor markets

What are you going to do with your data?

- What is the purpose of the primary data collection?
  - Describe a new phenomenon or new stylized facts
  - Identify a mechanism or estimate a parameter
  - Quantify a model
- Design the data collection in a way that enables this purpose

Example 1: The Value of Face-to-Face

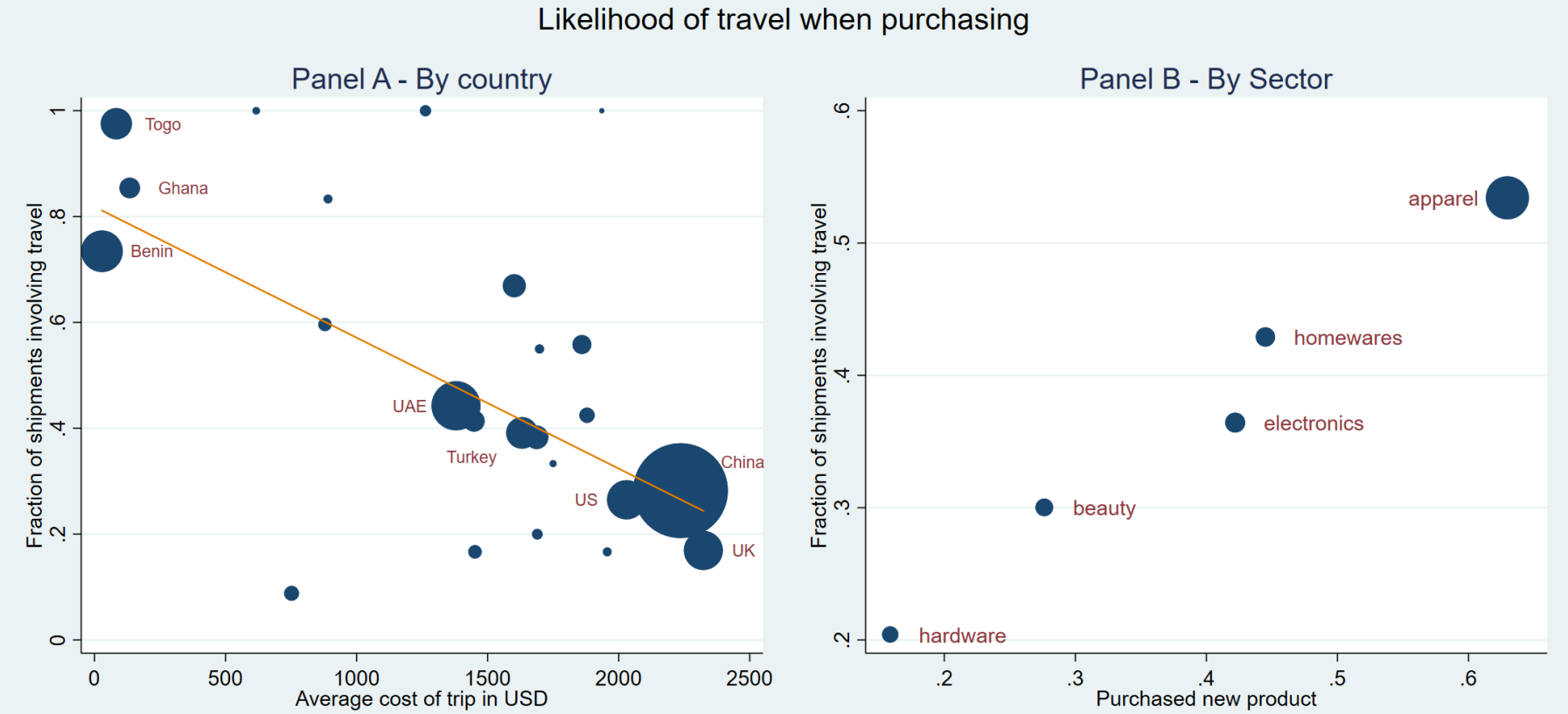
- The Value of Face-to-Face: Search and Contracting Problems in Nigerian Trade
  - Uses primary data from wholesale/retail traders in Lagos, Nigeria
  - Lagos Trader Survey (LTS) panel data

1. **Document new facts:** Traveling to buy from suppliers in person is a common strategy used by Nigerian traders
2. **Identify a mechanism:** This is a costly way to overcome product search and contract enforcement problems
3. **Quantification:** Estimates a model of choice to travel or order goods to quantify the size of those information frictions in trade

Data from the Lagos Trader Survey

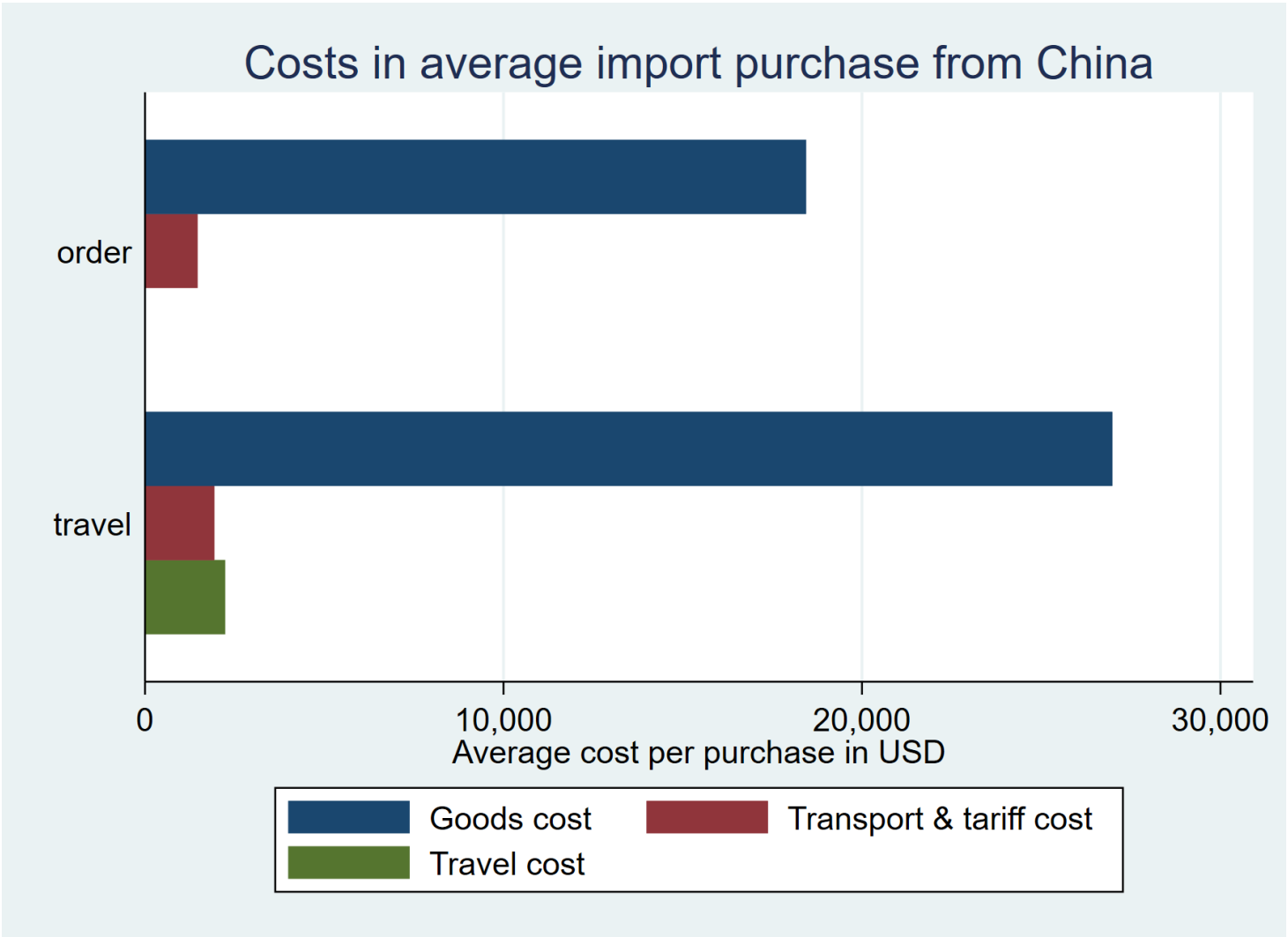
- Survey of 1,179 traders in Lagos
  - Sample from census
  - Panel of trader, shipment, and transaction-level info
- Deal in manufactured consumer goods
  - Apparel, electronics, toiletries, etc
  - Imported from China, UAE, US, Turkey, Benin, etc
- Small importers source for resale
  - Median 1 shop, 1 worker
  - Buy $67,000/year in imported stock

New facts I: Travel is a common part of sourcing



Note: Observations weighted by number of shipments

New facts II: Costly travel involves larger, less frequent transactions



Collect lots of details about new phenomenon

I'm going to ask you some details about some of the trips you personally took outside Nigeria for business in 2013 and 2014. Could we please start with the first trip you took IN 2013 OR 2014?

C.20 Where did you go on trip [number] in 2013 or 2014?	**C.21** When did you leave on this trip?	**C.22** How many days were you away on this trip?	**C.23** How many different suppliers did you buy from on this trip? **ENUM:** If they approached sellers in the open market and don't know who they were or how to find them again, write "1" here.	**C.24** On this trip, did you use paid agents for any of the following purposes? ***READ ALOUD and CHECK ALL THAT APPLY***	**C.25** Which of the following did you do on this trip? ***READ ALOUD and CHECK ALL THAT APPLY***
[country drop-down]	☐☐ month ☐☐☐☐ year	☐☐☐ days ☐ can't remember	☐☐ suppliers	1. ☐ To find products or suppliers 2. ☐ To translate 3. ☐ To arrange shipping 4. ☐ To inspect finished goods 5. ☐ None	1. ☐ Searched for new products or styles 2. ☐ Discussed a new design or product with a supplier 3. ☐ Inspected finished goods 4. ☐ Visited a supplier to maintain a relationship 5. ☐ Bargained over prices 6. ☐ None

C.26 How much money did you spend on **visas** for this trip to [*country name*] in [*month/year*]? **ENUM:** Enter -888 for "refused" or -777 for "don't know"	**C.27** How **long** did it take to receive the visa?	**C.28** How much money did you spend on **airfare** for this trip? **ENUM:** Enter -888 for "refused" or -777 for "don't know"	**C.29** How much money did you spend on **other travel expenses** for this trip, such as ground transportation, hotels, hiring agents, or money exchange? **ENUM:** Enter -888 for "refused" or -777 for "don't know"	**C.30** How much did you pay in total for transporting/**shipping** for ALL the products you bought on this trip to [*country name*]? **ENUM:** Enter -888 for "refused" or -777 for "don't know"	**C.31** How much did you pay in total **to clear the port in Nigeria** for everything you bought on this trip? Please include any tariffs, agent fees, and tips. **ENUM:** Enter -888 for "refused" or -777 for "don't know"	**C.32** Did you carry any products back with you in your luggage?	**C.33** What was the approximate value in total of all products you brought back in your luggage? **ENUM:** Enter -888 for "refused" or -777 for "don't know"
Currency: [drop-down] ☐☐☐,☐☐☐,☐☐☐ amount ☐☐☐,☐☐☐,☐☐☐ amount ☐ N/A – valid visa from a previous trip **>>> Skip to C.26** ☐ N/A – visa not required **>>> Skip to C.26**	☐☐☐ ☐ days ☐ weeks	Currency: [drop-down] ☐☐☐,☐☐☐,☐☐☐ ☐☐☐,☐☐☐,☐☐☐ ☐ N/A – did not travel by air ☐ Included in visa cost	Currency: [drop-down] ☐☐☐,☐☐☐,☐☐☐ ☐☐☐,☐☐☐,☐☐☐	Currency: [drop-down] ☐☐☐,☐☐☐,☐☐☐ ☐☐☐,☐☐☐,☐☐☐	☐☐☐,☐☐☐,☐☐☐ NGN ☐☐☐,☐☐☐,☐☐☐ NGN ☐ Included in transport/shipping cost	1. ☐ Yes 2. ☐ No **>>> Skip to C.32**	Currency: [drop-down] ☐☐☐,☐☐☐,☐☐☐ ☐☐☐,☐☐☐,☐☐☐

Data structure strategies

- Lots of descriptive detail
  - Not just whether they traveled, but details about costs and time and activities
  - Ask both quantitative and qualitative questions
  - Ask about specific trips and connect them to specific transactions
- Think very carefully about the data structure you generate
  - LTS surveys creates levels of detail within a single survey round
    - ★ E.g. a panel of past transactions: (a) Feb 2014: Men's jeans from Supplier Z on a trip (b) Feb 2014: Chinos from Supplier Y on a trip (c) June 2014: Chinos from Supplier Y ordered remotely (d) October 2014: Men's jeans from Supplier Y ordered remotely
  - Retrospective panel allows for mover-design type identification even from one data collection round
  - Can compare outcomes when traveling vs. ordering up to country-buyer-seller-product FEs

How to collect data about something new?

1. Serendipity and curious listening
   - Dinner next to a political scientist working on market associations

How to collect data about something new?

1. Serendipity and curious listening
   - Dinner next to a political scientist working on market associations
2. Exploratory fieldwork
   - Open-ended conversations with lots of traders

How to collect data about something new?

1. Serendipity and curious listening
   - Dinner next to a political scientist working on market associations
2. Exploratory fieldwork
   - Open-ended conversations with lots of traders
3. Piloting
   - Testing out structured questionnaires, validating qualitative hunch that this is important

How to collect data about something new?

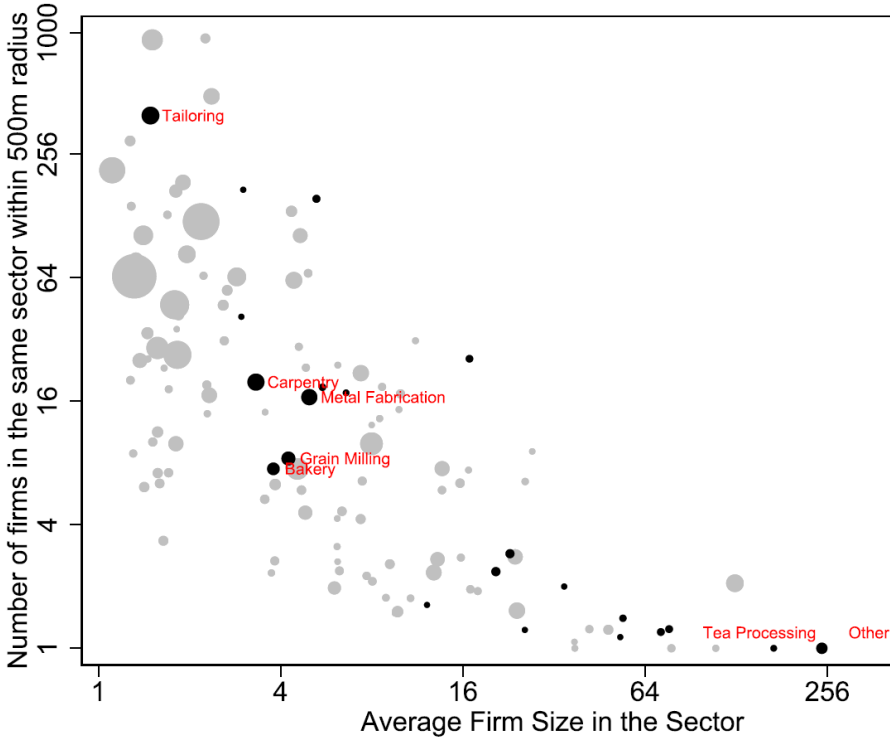
1. Serendipity and curious listening
   - Dinner next to a political scientist working on market associations
2. Exploratory fieldwork
   - Open-ended conversations with lots of traders
3. Piloting
   - Testing out structured questionnaires, validating qualitative hunch that this is important
4. Full survey
   - Actual "primary data collection" was the last step!

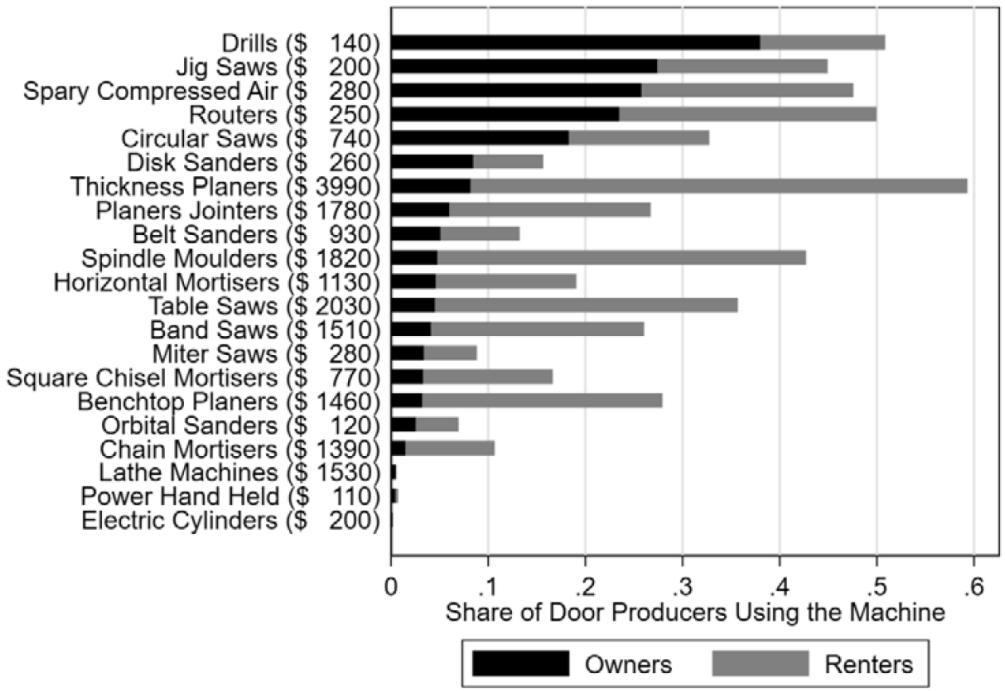
Example 2: Achieving Scale Collectively

- Achieving Scale Collectively (Bassi et al. 2022)
  - Uses primary data on manufacturing firms in 3 sectors in Uganda

1. **Document new facts:** Active rental markets for production equipment used by co-located manufacturing firms in Uganda
2. **Identify a mechanism:** This enables small firms to gain productivity advantages of mechanization without achieving the scale needed to purchase large capital equipment
3. **Quantification:** Estimates a model to quantify the impact of the rental market on productivity and firm size distribution

New facts: Lots of equipment rental for co-located firms





Incredible detail in survey instrument - 187 pages!

5.1c.def What's the model of this machine? I 99=Does not know	TEXT Q5_1c_def .....
5.1d.def How many %rostertitle% do you use in the production of %Q3_3c%?	NUMERIC: INTEGER Q5_1d_def -
5.2.def For how many hours is %rostertitle% operated per day by the firm in the production of %Q3_3c%? Please round to the closest integer V1 (self>0 && self<=24) || self==99 M1 The number of hours per day cannot exceed 24	NUMERIC: INTEGER Q5_2_def - SPECIAL VALUES 99 Does Not Know
5.3.def For how many days per week is %rostertitle% operated/running for the production of %Q3_3c%? V1 (self>0 && self<=7) || self==99 M1 The number of days per week cannot be more than 7	NUMERIC: INTEGER Q5_3_def - SPECIAL VALUES 99 Does Not Know
5.4.def Are the %rostertitle% that you are currently using property of the firm or are you renting them?	SINGLE-SELECT Q5_4_def 01 ○ machines are owned 02 ○ machines are rented 03 ○ some are rented some are owned

Estimating a parameter

- Model of firms' choice to rent vs. own
  - Likely to be transaction costs involved in renting, which are not fully observable/measurable

Estimating a parameter

- Model of firms' choice to rent vs. own
  - Likely to be transaction costs involved in renting, which are not fully observable/measurable
- Model provides an estimating equation:

$$log\left(\frac{\bar{p}_{sc}K_{sj}}{\bar{w}_j L_{sj}}\right) = \beta_0 + \beta_1 Rent_{sj} + \vartheta_s + \lambda_c + \delta X_j + \epsilon_{sjc}$$

- Identification problem: expect β_1 to be biased due to correlation between unobserved firm characteristics (in error term) and variables of interest

Estimating a parameter

- Model of firms' choice to rent vs. own
  - Likely to be transaction costs involved in renting, which are not fully observable/measurable
- Model provides an estimating equation:

$$log\left(\frac{\bar{p}_{sc}K_{sj}}{\bar{w}_j L_{sj}}\right) = \beta_0 + \beta_1 Rent_{sj} + \vartheta_s + \lambda_c + \delta X_j + \epsilon_{sjc}$$

- Identification problem: expect β_1 to be biased due to correlation between unobserved firm characteristics (in error term) and variables of interest
- Solution: Include firm fixed effects – β_1 now identified off of variation in utilization of rented vs. owned machines across production steps, WITHIN the same firm

Technical and practical issues – outline

1. Sampling
2. Survey design
3. Management
4. Data quality

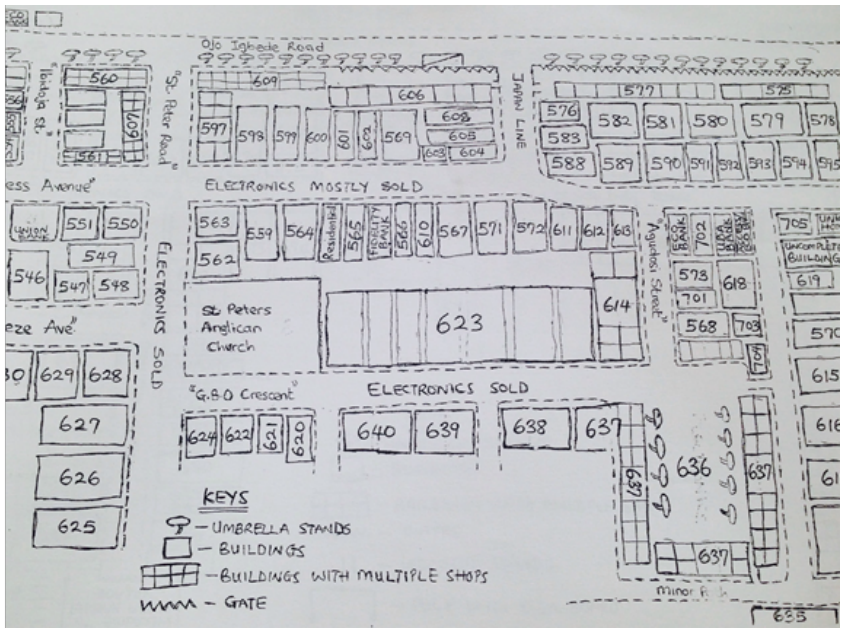
Sampling concepts

- Population of interest
  - Wholesale and retail firms in Nigeria?
  - Wholesale and retail firms operating from permanent physical premises in commercial areas of Lagos
- Sample frame
  - List of number of shops selling each product type on each floor of each building
- Sample
  - 1,200 shops – simple random sample from frame

Sample frame - Quasi-census of Lagos shops

Building a sample frame of Lagos wholesalers (in a mega-city of 25m)

- Started with a list of markets and plazas from the Lagos waste collection agency
- Census of buildings and shops in commercial districts on foot by enumerators
- 50,000+ shops enumerated by location & product



Sampling issues to consider

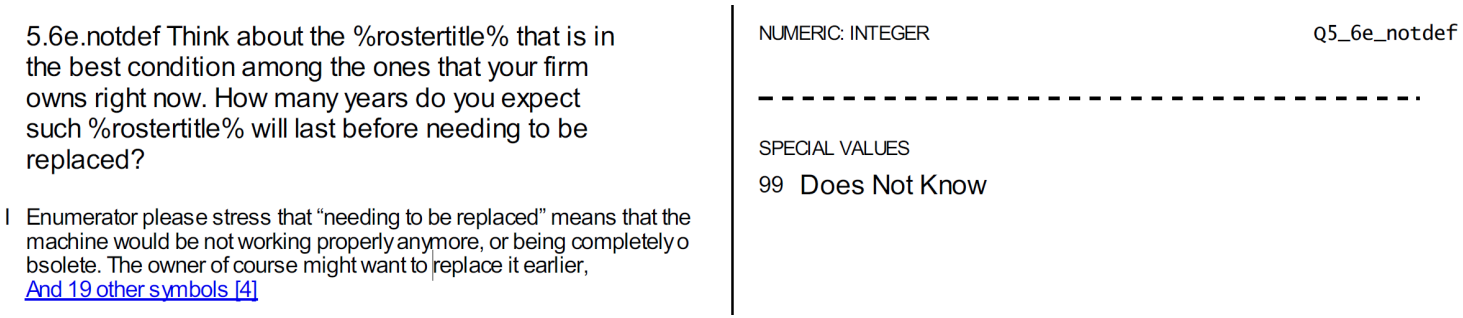
- How can you build a sample frame and track sampled units?
  - Common strategies: rosters or registration lists from local government, user lists from partner orgs, original listing/census
- Stratification and sampling weights, or clustered sampling
  - Ensure proportionality by subgroup
  - Oversample individuals of interest that are relatively rare (e.g. big firms)
- Sample size and power
  - What limits sample size? Budget, time, partner logistics/budget (for an RCT)
  - On the flip side, smaller sample usually means less statistical power in your analysis
  - Often good to run power calculations ahead of time to see what you can expect for different sample sizes.

Survey design - plan the big picture

- Who is the respondent?
  - Not always obvious based on sampling unit – e.g. which individual within a household or firm?
- How long can the survey be?
  - Varies a LOT from setting to setting
- What format will data collection take?
  - CAPI or paper (almost always the former now)
    - ★ What programming language / data collection software will you use?
  - Phone or in person or online?
  - Multiple or single visit?
  - Interview or enumerator observation?

Survey design – be pedantic!

- Think about the structure of the data that will come out, and how survey questions map to variables you will use in analysis
- Word questions in ways that are clear, specific, and comprehensible to the respondent



- Response options should
  - Be collectively exhaustive and mutually exclusive
  - Don't use categorical variables when you don't need to
  - Be clear about units and reference periods
- Use the power of CAPI to your advantage
  - Program in references and conditional paths through the survey
  - Randomize questions to shorten survey or test methodological issues

Management – who will collect the data?

- Survey firm, IPA/JPAL, local university/research institute, you?
  - Trade-offs and constraints around cost, local experience, experimental/non-exp methods, experience with academic researchers (vs. market research), extent of reputational incentives to do good work for you
- Ideally, work with an advisor or colleague who has done this before, in that country
- Screen your provider – look at detailed budgets, quality protection protocols, ask about enumerator selection (language, experience, local, education?) and training

Management - ethics, privacy, data control

- IRB – institutional review board
  - In the US, standards are based on a very specific history, largely designed to apply to medical research
  - Depending on the setting, there may (or may not) also be local IRB-type procedures to follow
- Informed consent
  - Think about language, literacy, background knowledge of population
  - May need to provide training and oversight to make sure this is done appropriately
- Data security
  - Who has access to personally identifying info?
  - Think about weak points in the chain (often devices used by enums, or email with partner orgs)
- Pre-analysis plan or registration

Quality – survey rollout

- Be there in person for piloting and training whenever possible
- Roll out a survey in stages
  - Exploratory interviews
  - Piloting
  - Soft launch
  - Full launch
- Go slowly!
  - Look at the data, revise, and iterate based on what you learn at every step
  - You will get pushback – everyone else's incentive is to go as quickly as possible
  - Limit survey completion rates and build in breaks to evaluate and think

Quality – quality control processes

- LOOK AT YOUR DATA. LOOK AT YOUR DATA. LOOK AT YOUR DATA. (Before it's too late.)

Quality – quality control processes

- LOOK AT YOUR DATA. LOOK AT YOUR DATA. LOOK AT YOUR DATA. (Before it's too late.)
- Audits
  https://www.povertyactionlab.org/resource/data-quality-checks
- High frequency checks
  https://dimewiki.worldbank.org/High_Frequency_Checks
- Spot checks, sit-ins with enumerators
- Build in internal sanity checks – e.g. ask the same question multiple ways, flag a problem to make the enumerator go back if things don't match across sections of the survey or are implausibly high/low values

Big picture strategy for primary data collection

- Be clear on what you want to get out of the data collection
- Being exploratory is a great thing to do -- use (small scale, open-ended) data collection to figure out how things actually work and learn what questions need answering
- If you want to use the data for quantitative analysis or estimation
  - Have a sketch of a model or a research design
    - ★ To know what parameters to identify or variables to measure
  - Think about where identification is going to come from, and how much variation and power you will have
    - ★ Across individuals? Sectors? Markets? Villages? Time? Treatment/control groups?)
  - Think about the structure of the data that will come out
    - ★ Unit of observation, time period is covered, panel or multiple observations on a non-time dimension per unit

Top tips

- Go deep, not wide on the topics you cover
- Be creative about how to get people to answer questions
- Ask quantitative questions, not just categorical (e.g. ask what price they charge, not whether they raised or lowered their price)
- Ask different questions – there's no point in asking the same questions everyone else does
- Don't be too confident in your priors – design your data collection so that it is possible to learn if the way you are thinking about things is off-base

Resources

- DIME Wiki (World Bank): https://dimewiki.worldbank.org
- JPAL Research Resources: https://www.povertyactionlab.org/research-resources?view=toc#choose-a-view
- World Bank Development Impact blog: https://blogs.worldbank.org/en/impactevaluations
- SurveyCTO: https://www.surveycto.com/