Module 3: Inclusive Data Collection Methods

Inclusion Data Quest

WE ARE OPEN

Module Objective

This Module explores inclusive data collection methods. We will cover survey design, data sources, and technology.

We will also discuss stakeholder collaboration. Learn to blend institutional & grassroots data.

Agenda

1. Data Source Selection
2. Survey design for inclusion
3. Technology for Data Collection
4. Stakeholder Collaboration - for inclusive data
5. Capacity Building
6. Monitoring & Evaluation
7. Key takeaways

1. Data Source Selection

Deciding on source of data

1. **Define Objectives** — Clarify research questions. **What?**
2. **Evaluate Sources** — Assess reliability and bias. **Why?**
3. **Combine Sources** — Triangulate findings. **How?**

Primary vs Secondary data

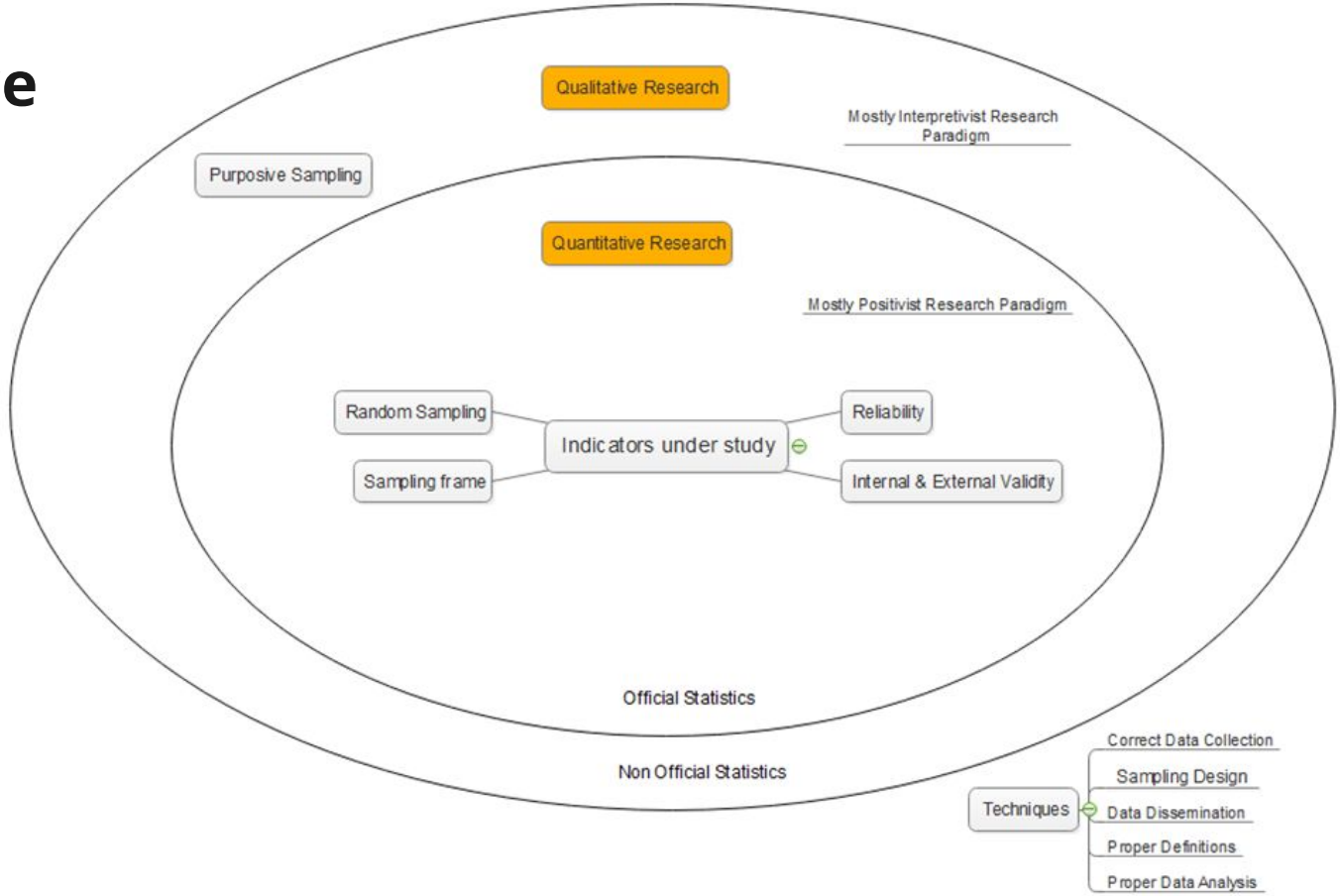
In the world of research and data analysis, two main types of data are used: these data can be classified into two categories, namely primary data and secondary data.

1. **Primary Data**:
   a. Collected firsthand by the researcher for a specific study or purpose.
   b. *Example*: Survey responses, interviews, lab experiments.
2. **Secondary Data**:
   a. Pre-existing data gathered by others for a different purpose.
   b. *Example*: Government reports, academic journals, historical archives

Criteria	**Primary**	**Secondary**
Source	Collected first hand by researcher	Collected by others
Time and Cost	Takes longer; more expensive	Faster and Cheaper
Controls	Researcher controls variables, design (sample/methodology	Limited control

‘HOW’ is more important than ‘WHO’

‘How’ data is being collected to answer the needs at hand or ‘who’ does the data collection as long as rigorous methodology is adhered to and communicated



Do we do a survey or rely on administrative data?

Survey data refers to when a particular phenomenon is under investigation – requiring responses from individuals or households. This requires more rigor in terms of validity/reliability tests; hence sampling is a critical consideration when it comes to survey data so that the findings can be generalized to the population

Survey data *- would require more rigor in terms of validity/reliability tests; for example sampling is very important (in Kenya for example theres is the NASSEP frame which is very important for surveys.)*

Administrative data refers to daily operational data collected by County Departments on the state of affairs, such as roads, disease burden, enrolment. As a natural count of ALL cases of the phenomenon, this data often might not require rigorous validity and reliability tests.

Administrative data *– roads, disease burden, enrolment, does not need a lot of validity tests (it can be sampled randomly to ensure that it is representative of the population).*

New waves of data

- Alternative data sources such as **citizen-generated data (CGD), mobile data**, **geospatial data**, and **big data** have become increasingly relevant.
- They are defined as follows by the World Data Report (2021):
  - **CGD**: are produced by individuals, often to fill gaps in public and private sector data or when the accuracy of existing data is in question
  - **Geospatial Data**: Public intent geospatial data include satellite imagery of the Earth such as NASA data or weather data
  - **Big data**: data sets that are difficult to store and process using common software tools, regardless of the computing power or the physical storage at hand; for example social media, etc.
  - **Other non-traditional sources**: for example from crowdsourced platforms (like, Ushahidi) to capture informal or transient populations

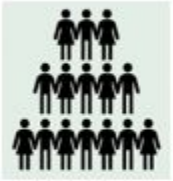
Summary: Sources of data

Administrative data

are generated by a process of registration or record keeping, usually by national authorities such as birth, marriage, and death records, tax records etc.

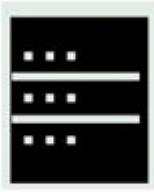
Citizen generated data

are produced by individuals, often to fill gaps in public and private sector data or when the accuracy of existing data is in question



Census data

aim to systematically enumerate and record information about an entire population of interest, whether individuals, businesses, farms, or others



Machine generated data

are automatically generated by a sensor, application, or computer process without human interactions. An example is the sensors that monitor air pollution

Sample survey data

draw on a smaller, representative sample of the entire population, typically from censuses, to collect detailed information more frequently

Geospatial data

Public intent geospatial data include satellite imagery of the Earth such as NASA data or weather data

2. Survey design for inclusion

Quantitative vs Qualitative Data

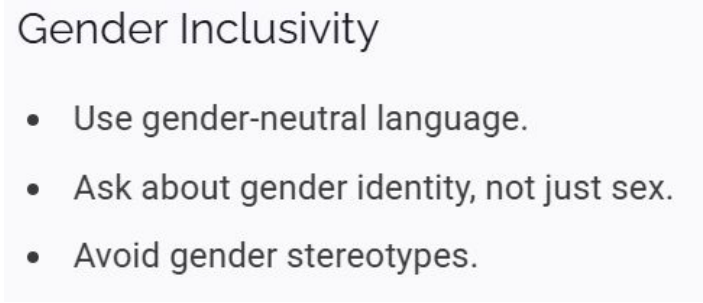
Deciding what data is required lays the right foundation on the kinds of decisions, analysis and conclusions that can subsequently be made.

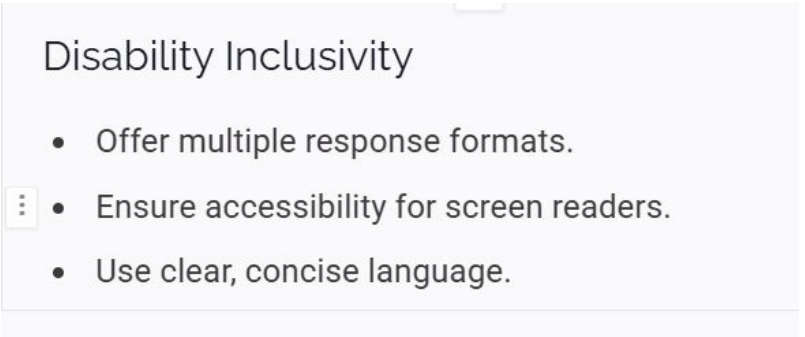
Quantitative	***Qualitative***
Numbers, information that can be counted.	*Information about behavior, perception. It can seldom be counted*

Quantitative	***Qualitative***
Examples	
• *How many people took part?* • *How much did it cost?* • *How long was it?* • *What happens when, who it is for, cost, etc.*	• *Sharing likes and dislikes* • *How they think it could be improved* • *What difference it has made to their lives* • *Whether they would recommend the programme to others*
Data Sources	
• *Statistics and statutory data* • *Structured surveys involving closed questions, multiple choice or likert scale items* • *Positivist research paradigm*	• *Stories of participants' experiences and impact* • *Observation* • *Open ended interviews* • *Interpretivist research paradigm*

Tool/Questionnaire Design for Inclusion

- **Cultural Relevance**: Avoid assumptions or stereotypes, recognize diversity (e.g., gender-neutral options for marital status). Use local dialects and context-specific examples.
- **Accessibility**: Design tools that work with assistive technologies (e.g., screen readers for visually impaired users) and use plain language to avoid jargon.
- **Multiple Formats**: Offer audio, visual, or tactile interfaces (e.g., Braille surveys, pictorial scales for low-literacy populations).
- **Pilot Testing**: Test tools with diverse groups (e.g., people with disabilities, ethnic minorities) to identify exclusionary design flaws.
- Create new approaches to understanding the lived – varied, innovative and flexible





What is Pretesting?

Pretesting refers to the testing of the questionnaire on a small sample of respondents to identify and eliminate potential problems

Why does it matter?

- A questionnaire should not be used in the field without adequate pretesting
- All aspects of the questionnaire should be tested, including question content, wording, sequence, form and layout, question difficulty and instructions
- Respondents for the pretest and for the actual survey should be drawn from the same population
- Pretests are best done by personal interviews even if the actual survey is to be conducted by mail, telephone or electronic means because the interviewers can observer respondents reactions and attitudes

2 Research Paradigms (relevant to this course)

Positivism

- Assumes an objective reality that can be measured and quantified.
- Relies on empirical data, experiments, and statistical analysis.
- Prioritizes generalizability and replicability.
- **Inclusive Consideration**: Risks excluding subjective experiences of marginalized groups if overly rigid.

Interpretivism

- Focuses on subjective meanings and social contexts.
- Uses qualitative methods (e.g., interviews, ethnography).
- **Inclusive Strength**: Captures lived experiences of underrepresented communities.

Differences between positivism and interpretivism

Approach

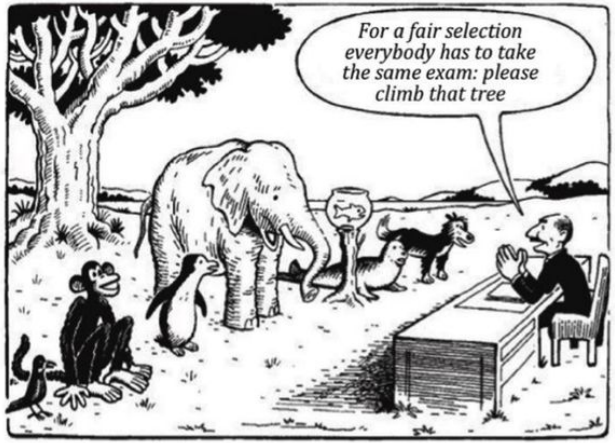
Positivism: Objective and measurable | Interpretivism: Subjective and interpretive

Data collection

Positivism: Quantitative methods | Interpretivism: Qualitative methods

Researcher's role

Positivism: Objective observer | Interpretivism: Active participant



Our Education System

"Everybody is a genius. But if you judge a fish by its ability to climb a tree, it will live its whole life believing that it is stupid."

- Albert Einstein

3. Technology for data collection

Technology is a tool

- Technology is an enabler for efficiency and effectiveness
- Data management is the creation and implementation of **architectures**, **policies**, and **procedures** that manage the **full data lifecycle** needs of an organization.

Beginners Pitfalls

- In the absence of a coherent strategy, data quality can suffer. The 'garbage in, garbage out' is particularly apt here. Without established processes for data collection and normalization, the data gathered can be of poor quality, leading to misguided decisions and strategies
- Poor data practices don't just affect internal operations; they can have far-reaching consequences.
- In the realm of public data and **governance**, this could mean **misinformed policies, inefficient public services, and a loss of public trust.**

Effective data stack

At its core, an effective data stack makes it possible to perform six basic operations:

1. **Collection**: This is the process of gathering data from various sources e.g data from the field, excel sheets, web and mobile tools, etc. The key here is not just the quantity of data, but its **relevance** and **quality**.
2. **Normalization**: Before being stored, data often needs to be formatted, combined, or normalized. Here, the focus is on **standardizing** and **cleaning** the data.
3. **Storage**: Once ingested, data needs to be stored in a place where it’s accessible for analysis. This isn't just about having enough space; it's about ensuring data **security**, **privacy**, and **accessibility**.

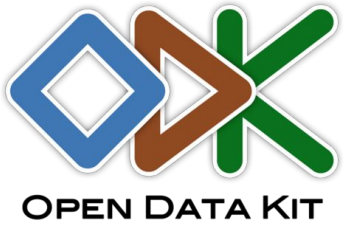
Effective data stack

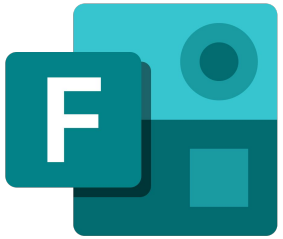
4. **Transformation**: Once stored, raw data is prepared for analysis. Common transformations include joining tables, creating aggregations, and building in key business logic. Ensures that when we **analyze** the data, we can extract **meaningful** and **actionable** insights.

5. **Optimization**: As datasets grow, further considerations need to be made—can the data be formatted to optimize the space it takes up, or the time it takes to query? In this stage, the focus is on improving the **efficiency** and **performance** of the data stack.

6. **Analysis**: This is where data is turned into insights. Analysis can range from simple queries and visualizations to complex predictive modeling and machine learning. The goal here is to **derive** actionable **intelligence** that can inform decisions, policies, and strategies.

Examples of (mobile)tools for data collection





6 critical features of mobile data collection tools

- **Offline Functionality:** Ability to collect and store data without internet connectivity, syncing later when a connection is available.
- **Multi-Media Support**: Capture photos, videos, audio recordings, barcodes, or GPS coordinates alongside text/numeric data
- **Real-Time Data Sync & Cloud Integration**: Automatically upload data to cloud platforms (e.g., Google Drive, AWS) for instant access and collaboration.
- **Form Logic & Skip Patterns**: Dynamic forms that adapt based on user responses (e.g., hiding irrelevant questions).
- **Data Validation Rules:** Built-in checks (e.g., range limits, mandatory fields, format constraints) to ensure data quality.
- **Cross-Platform Compatibility**: Works seamlessly on Android, iOS, and web browsers with consistent functionality.

Reflection

What are your key takeaways from this discussion?