HOW TO COLLECT GOOD QUALITY DATA

DR. SAMUEL OTI
Winner of United Nations Population Award
2015

A global center of excellence, consistently generating and delivering relevant scientific evidence for policy and action in Africa
What we do

RESEARCH
GENERATION AND SYNTHESIS OF SCIENTIFIC KNOWLEDGE

AGING AND DEVELOPMENT
EDUCATION RESEARCH
HEALTH CHALLENGES & SYSTEMS
POPULATION DYNAMICS & REPRODUCTIVE HEALTH
URBANIZATION & WELLBEING

STATISTICS AND SURVEYS

POLICY ENGAGEMENT & COMMUNICATIONS

- Communications
- Policy Engagement
- Knowledge Management

RESEARCH CAPACITY STRENGTHENING

- Training Programs
- Partnership with Universities
- Fellowship Programs
15 YEARS....AND COUNTING...

- Nearly 450 Peer reviewed journal articles
- Over 100 Technical reports and policy briefing

- 270 PhD fellows
- 40 Post-Doctoral fellows
Kindly memorize this econometric model for data quality...

\[ \sum_{i=0}^{\infty} (\lambda_1 \theta)^i E_t x_{t+i} = \sum_{i=0}^{\infty} (\lambda_1 \theta)^i \left[ \frac{\sigma(L)}{L^i} \right] + \frac{1}{\sigma(L)} x_t = \]

\[ = \left\{ \left( 1 + \sigma_1 L + \sigma_2 L^2 + \cdots \right) + (\lambda_1 \theta) L^{-1} \left( \sigma_1 L + \sigma_2 L^2 + \sigma_3 L^3 + \cdots \right) + \right. \]

\[ + (\lambda_1 \theta)^2 L^{-2} \left( \sigma_2 L^2 + \sigma_3 L^3 + \sigma_4 L^4 + \cdots \right) + \cdots \right\} x_t / \sigma(L) = \]

\[ = \left\{ \left( 1 + \sigma_1 L + \sigma_2 L^2 + \cdots \right) + (\lambda_1 \theta) L^{-1} \left( 1 + \sigma_1 L + \sigma_2 L^2 + \cdots \right) \right. \]

\[ + (\lambda_1 \theta)^2 L^{-2} \left( 1 + \sigma_1 L + \sigma_2 L^2 + \cdots \right) + \cdots - \left[ (\lambda_1 \theta) L^{-1} + \right. \]

\[ + (1 + \sigma_1 L)(\lambda_1 \theta)^2 L^{-2} + (1 + \sigma_1 L + \sigma_2 L^2)(\lambda_1 \theta)^3 L^{-3} + \cdots \left] \right\} \rho(L) x_t = \]

\[ = \left\{ \sigma(L) \left[ 1 + (\lambda_1 \theta) L^{-1} + (\lambda_1 \theta)^2 L^{-2} + \cdots \right] - \right. \]

\[ - (\lambda_1 \theta) L^{-1} \left[ 1 + (1 + \sigma_1 L)(\lambda_1 \theta) L^{-1} + (1 + \sigma_1 L + \sigma_2 L^2)(\lambda_1 \theta)^2 L^{-2} + \right. \]

\[ + (1 + \sigma_1 L + \sigma_2 L^2 + \sigma_3 L^3)(\lambda_1 \theta)^3 L^{-3} + \cdots \right\} \rho(L) x_t \]
Outline

• Case study – The **Tayari** Project
• Data management process
  – What could go wrong?
• Ensuring data quality
  – Intrinsic
    • Design and testing instruments
    • Recruitment and training of staff
    • Field procedures
    • Data entry, cleaning & storage
    • Data analysis
  – Extrinsic
The Tayari Project

- ECE intervention targeting **preschools** in public and low cost private centers in Kenya
- Study location: 3/47 counties in Kenya – **Kisumu, Machakos and Nairobi**
- Duration: 2015 - 2018
- Implementers: RTI and CIFF
- Evaluators: APHRC
The Tayari Project

Intervention components:

I. A classroom instruction model (student materials, teacher guides, instructional training and ongoing coaching support) that delivers quality learning for children aged 3 – 6

II. A package involving the classroom instruction model + a nutrition and health intervention. A psychosocial component will be integrated in these interventions.
### The Tayari Project

**Intervention implementation: Phased-in**

<table>
<thead>
<tr>
<th>Arm</th>
<th>2015</th>
<th>2016</th>
<th>2017</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classroom instruction only (N=600)</strong></td>
<td>150 centres (90 public + 60 LCPC)</td>
<td>210 centres (150 public + 60 LCPC)</td>
<td>240 centres (180 public + 60 LCPC)</td>
</tr>
<tr>
<td><strong>Classroom instruction + a nutrition and health intervention (N=105)</strong></td>
<td>--</td>
<td>45 public centres</td>
<td>60 public centres</td>
</tr>
</tbody>
</table>
The Tayari Project – Impact Evaluation

Three Counties: Kisumu, Machakos & Nairobi

Kisumu
Control (public & LPCP)  Treatment 1 (public & LPCP)  Treatment 2 (public)

Machakos
Control (public & LPCP)  Treatment 1 (public & LPCP)  Treatment 2 (public)

Nairobi
Control (public & LPCP)  Treatment 1 (public & LPCP)  Treatment 2 (public)

From selected zones within each county, we randomly sample preschools.

Final Evaluation Sample

Control (65 in 2015)
Treatment 1 (65 in 2015 + 20% of 2016 + 20% of 2016)
Treatment 2 (45 in 2016 + 20 in 2017)

Notes:
Treatment 1 = Classroom instruction
Treatment 2 = Classroom instruction + health & nutrition
The Tayari project – IE Data collection

• **Quantitative:**
  - Head teacher questionnaire
  - ECE teacher questionnaire
  - Kenya School Readiness Assessment Tool (KSRAT)
  - Classroom lesson observation schedule

• **Qualitative:**
  - Key Informant Interviews
  - In Depth Interviews
  - Focus Group Discussions
Data management process

- Selection bias
- Manipulation of results
- Loss or theft of data
- Data entry errors
- High non-response rate
- Wrongful inclusion or exclusion of variables

WHAT COULD GO WRONG??

- Response or recall bias
- Fabrication of results

30/07/2015
Dr. Samuel Oti
All you need to know about data quality...

“Garbage in, garbage out”

--(Not sure who said this)
Ensuring data quality
Intrinsic

- Design and testing instruments
- Recruitment and training of staff
- Field procedures
- Data entry, cleaning & storage
- Data analysis
Design and testing instruments

- Allocate sufficient time to carefully develop instrument
- Add additional time if instrument is to be programmed electronically
- Search literature for standardized and validated instruments
- If not, make plans to **PRE-TEST** non-validated instruments
- Translate to local languages (Why?)
Recruitment and training of staff

- Develop clear job descriptions
- Recruit qualified persons (avoid overqualified persons! Why?)
- As much as possible, recruit from local community (why is this important?)
Dear Hubby,

I’ll be home a bit late tonight, please peel half of the potatoes on the table and put them to boil. I’ll finish the cooking when I get home.

Love,
Wifey
Last night I asked my husband to put some spaghetti on the stove so I could start dinner when I got home. I came home to this...
Recruitment and **training** of staff

• Typical contents of training (1)
  – Pre evaluation
  – Aims and objectives of project
  – Explain roles and responsibilities
  – Research ethics
  – **Question by question (all versions of instruments)**
Recruitment and training of staff

- **Question by question**
  - Ensures understanding of meaning and intent behind each question
  - Checks for errors in text, language and meaning
  - Ensures skip patterns are correct
  - Flags any inappropriate content
Recruitment and training of staff

• Typical contents of training (2)
  – Field logistics
    • Reaching and identifying participants
    • How to handle non-response, refusals
    • Admin (finance, transport, security, code of conduct)
  – Role plays
  – Post evaluation
  – Special session for supervisors (their roles and responsibilities)
  – Fake launch or pilot test and feedback
Supplementary materials

- Manuals and SOPs
- Informed consent forms
- Training guides
- Code-sheets
- Referral slips
- Vouchers
- Reporting forms
Field procedures

- Sit-ins
- Spot-checks
- Office editing
- Field corrections
- Data transfer process
Data entry, cleaning & storage

Data entry:
- Double data entry is ideal if paper-based
- Otherwise check 5-10% of captured data frequently
- Electronic not as error prone but must also be checked
- Start data entry as early as possible (should not be after data collection is over. Why?)
Data entry, cleaning & storage

Data cleaning:
- Work closely with data analyst if not doing it yourself
- How to handle missing values
- Implement data documentation protocol
- Data ‘locking’ & storage
Son, I don't see your name here...

Can you please check under the 'Publish or Perish' category?!
Data analysis

- Develop a **pre-analysis plan**
- **Register** the plan
- Stick to the plan and report any additional analysis
- Aim to publish results whether **negative or positive, expected or unexpected**
Extrinsic

• Community mobilization
• Involve local authorities
• Get necessary approvals
• Unforeseen external challenges
  – Security threats
  – Interference by local politicians or leaders
  – High refusal rates
Bottom-line

As much as possible prepare for and deal with data quality issues before and during the study. Post-hoc statistical manipulations may help but cannot perform miracles if the data is of poor quality to begin with...in other words BE PREPARED!
Questions?
Contacts:

Dr. Samuel Oti
African Population and Health Research Center
APHRC Campus, 2nd floor, Manga Close, Off Kirawa Rd
Nairobi-Kenya
Tel: +254 20 400 1058 Mobile: +254 738 515881
Email: soti@aphrc.org or dr.otisamuel@gmail.com
Web: www.aphrc.org