Exporting CAPI data to Stata: Experience from *surveybe*

Joachim De Weerdt
Research Director
EDI (Economic Development Initiatives)
Introduction

• A lot of data analysed in Stata are collected through surveys.
• Movement away from paper questionnaires, in favour of electronic data collection:
  1. Web surveys
  2. Phone surveys,
  3. Computer Assisted Personal Interviewing (CAPI), etc.
• Web and phone surveys are small
  – Few dozen variables, easy to manually clean and label
  – Typically one flat file: no problems of relational integrity
• But CAPI is now handling long, complex, multi-topic questionnaires:
  – Hundreds of variables, very hard to manually clean and label
  – Dozens of files all linked together: ensure relational integrity
Why a surge in CAPI

• Market developments:
  – Tablets are becoming cheaper (need a tablet for larger screen)
  – Mobile phone networks can handle data transfer from the field and are prolific, even in rural areas in developing countries
  – Software options are becoming available (e.g. surveybe)

• Compelling list of benefits:
  – Errors are identified and resolved during interview
  – One-click export to Stata in labelled data set
  – Data available immediately
  – Lower costs
  – Exponentially expands level of creativity: do stuff you cannot do on paper (call soundex algorithms, use regular expressions, access large look-up tables during interviews)
Reduce Errors:

We expect error to be reduced by:

• Automated routing (skips)
• Pre-coded drop down menus or radio-buttons
• Automatically captured GPS coordinates and interview times
• Elaborate system of validation checks
• Complex reports
Clean and labelled data set

- Variable labels: question text
- Value labels: drop-down text
- Complete relational integrity (_m==3!!)
Data Quality Matters

<table>
<thead>
<tr>
<th></th>
<th>(1) CAPI</th>
<th>(3) Pen and Paper</th>
<th>Significance of $t$-test that $(1) = (3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poverty Headcount</td>
<td>83.0</td>
<td>68.3</td>
<td>***</td>
</tr>
<tr>
<td>Gini (95% CI)</td>
<td>.24 (.22-.25)</td>
<td>.30 (.27-.32)</td>
<td>***</td>
</tr>
<tr>
<td>$N$</td>
<td>1200</td>
<td>319</td>
<td></td>
</tr>
</tbody>
</table>

From Caeyers, Chalmers and De Weerdt (2010)

In a randomised experiment we see that inequality goes up when consumption expenditures are measured on paper; measurement error is being picked up as income inequality.
Testing for attenuation bias

<table>
<thead>
<tr>
<th></th>
<th>No. of years of schooling (children aged 7-14)</th>
<th>Schooling expenditures on school-going children</th>
<th>Child slept under a treated bednet night before survey</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Age FE</td>
<td>Age FE</td>
<td>LPM Age FE</td>
</tr>
<tr>
<td>Log total consumption per aeu</td>
<td>0.204 (0.135)</td>
<td>7,775*** (1,630)</td>
<td>0.134*** (0.036)</td>
</tr>
<tr>
<td>( \delta )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>log total consumption per aeu* CAPI assignment</td>
<td>0.292** (0.144)</td>
<td>6,279*** (1,722)</td>
<td>0.154*** (0.040)</td>
</tr>
<tr>
<td>( N )</td>
<td>2,683</td>
<td>2,137</td>
<td>5,148</td>
</tr>
</tbody>
</table>

From Caeyers et. al (2010)
We see evidence of attenuation bias when using paper questionnaires: coefficients estimated with data collected on paper are biased towards zero.
SURVEYBE DEMO
Food for Thought

• Using different scripts and different languages
• Exporting the system of skip & validation rules
• Comments