Data Entry: Tips & Tricks
So you want to run a household survey...

Step 1: contract data collection agency
Step 2: design questionnaire
Step 3: pilot questionnaire
Step 4: draft survey manual
Step 5: train enumerators
Step 6: plan field work
Step 7: monitor data collection
Step 8: design data entry system
Step 9: supervise data entry
Step 10: data management and cleaning
So you want to run a household survey...

Step 1: contract data collection agency
Step 2: design questionnaire
Step 3: pilot questionnaire
Step 4: draft survey manual
Step 5: train enumerators
Step 6: plan field work
Step 7: monitor data collection
Step 8: design data entry system
Step 9: supervise data entry
Step 10: data management and cleaning
Double data entry process

- Enter each questionnaire twice (two different data entry clerks)
- Compare first and second entry to identify data entry errors (usually in stata)
- Check discrepancies by referencing hard copy
  - Physical questionnaires must be accessible and clearly catalogued – sort
- Make corrections to produce ‘reconciled’ data
FC role

- Make sure TORs are very precise
- Work closely with firm to set up data entry system **before field work starts**
  - confirm software meets specifications in TOR
  - provide data dictionary
- Check data entry system exhaustively **before data entry starts**
FC role

- Make sure TORs are very precise
- Work closely with firm on set up data entry system **before field work starts**
  - confirm software meets specifications in TOR
  - provide data dictionary
- Check data entry system exhaustively **before data entry starts**
TORs

- **Requirements for data entry template**
  - Looks like the questionnaire
  - Logic mimics flow of questionnaire
  - Range; variable type; consistency checks in real-time
  - Directly export fully-labeled data to stata
- **Double data entry a must**
  - different data entry clerks for each entry ("blind")
  - Full double data entry of all instruments
  - "reconciliation" of identified errors from hard copies
TORs

- delivery of **RAW** data
  - Must get raw data on ongoing basis
  - Very important that this never be modified!
- delivery of reconciled data
  - All errors identified in reconciliation process fixed
- payment conditional on acceptable error rate
  - Each batch of data tested by FC
FC role

- Make sure TORs are very precise
- Work closely with firm on set up data entry system **before field work starts**
  - confirm software meets specifications in TOR
  - provide data dictionary
- Check data entry system exhaustively **before data entry starts**
Software options

- Microsoft Excel
  - Common, but very problematic!
- CSPro
  - High functionality AND free
- Licensed softwares [Microsoft Access, SPSS]
  - High functionality BUT can be expensive
- Customized program written in visual basic
  - High functionality BUT need to have programmer
Data entry software

- Data dictionary is backbone of data entry template. Best if FCs create from questionnaire

- Components
  - Variable name
    - = question number for easy reference to questionnaire
    - Start with letter of module (Section B, question 10 → B10)
  - Variable label
    - = text of question
    - paraphrase if more than 80 characters
  - Value set
    - = all possible values for the variable
    - Categorical: all categories, with labels
    - Continuous: range
    - String: set length (number of characters allowed)
Other requirements
- Consistent codes for missing and skipped fields
- Multiple choice variables converted to unique binary variables in data
- Mask includes date, time and data entry clerk ID
- Built-in checks to make sure IDs are unique and within sample frame

Software matters, but even more important is close interaction with the programmer!
Good plan for physical storage of questionnaires is key

- 3,000 40-page surveys take up a LOT of space!
- Need to be clearly catalogued by cluster and unique ID for easy reference
  - Critical for reconciliation process
- Make sure TORs are very precise
- Work closely with firm on set up data entry system **before field work starts**
  - confirm software meets specifications in TOR
  - provide data dictionary
- Test data entry system exhaustively **before data entry starts**
Testing data entry software

- Critically important!!
- What to check:
  - Flow and skip patterns
  - Values and ranges
- Enter at least a dozen questionnaires in full
  - Use pilot data
  - Keep track of how long it takes to enter 1 questionnaire
- Export data to stata and check output
  - labels are maintained and visible in full?
  - Values appear correctly, no formatting problems?
  - Send output to research team for confirmation
So you want to run a household survey...

Step 1: contract data collection agency
Step 2: design questionnaire
Step 3: pilot questionnaire
Step 4: draft survey manual
Step 5: train enumerators
Step 6: plan field work
Step 7: monitor data collection
Step 8: design data entry system
Step 9: supervise data entry
Step 10: data management and cleaning
FC role

- Help train data entry operators
- Ongoing supervision of data entry process
- Check the data!!
  - Start early! Do not wait until data entry is finished!
- Check number of entered questionnaires for each cluster compared to survey log
FC role

- Help train data entry operators
- Ongoing supervision of data entry process
- Check the data!!
  - Start early! Do not wait until data entry is finished!
- Check number of entered questionnaires for each cluster compared to survey log
Train data entry operators

- Lead training together with data manager
- Make sure flow of questionnaire is well-understood (especially rosters and tables)
- Explain what the project’s about
  - Data entry can be boring! Quality of work is better if DEOs are motivated
- Data entry clerks should practice at least two days before entering “real” data
- Be available to answer questions
FC role

- Help train data entry operators
- **Ongoing supervision of data entry process**
- Check the data!!
  - Start early! Do not wait until data entry is finished!
- Check number of entered questionnaires for each cluster compared to survey log
Ongoing data entry supervision

- Verify system for backing up data
  - Daily backups on each machine
  - Weekly backups on external hard drive
- Drop-in unexpectedly to verify reconciliation process is going on as planned
- Troubleshoot field errors
  - Data entry team should save all questionnaires with problems aside for you to check
  - Do not encourage data entry team to do their own problem solving!
FC role

- Help train data entry operators
- Ongoing supervision of data entry process
- Check the data!!
  - Start early! Do not wait until data entry is finished!
- Check number of entered questionnaires for each cluster compared to survey log
Check data

- Survey firm (usually) does double-entry comparison, but FC must re-do and confirm using raw data
- Stata ado files make this very easy
  - c fout
  - Readreplace
  - used to be much more complicated – thanks IPA!
Example first and second entry files

<table>
<thead>
<tr>
<th>variable name</th>
<th>storage</th>
<th>type</th>
<th>display format</th>
<th>value label</th>
<th>variable label</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>byte</td>
<td>%8.0g</td>
<td>unique id</td>
<td></td>
<td></td>
</tr>
<tr>
<td>vlg_code</td>
<td>byte</td>
<td>%8.0g</td>
<td>village code</td>
<td></td>
<td></td>
</tr>
<tr>
<td>maize_harvest</td>
<td>int</td>
<td>%8.0g</td>
<td>quantity of maize harvested in 2012 in kgs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hhh_age</td>
<td>byte</td>
<td>%8.0g</td>
<td>age of HH head</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### First entry

<table>
<thead>
<tr>
<th>id</th>
<th>vlg_code</th>
<th>maize_harvest</th>
<th>hhh_age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>42</td>
<td>1500</td>
<td>55</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1900</td>
<td>48</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>1800</td>
<td>22</td>
</tr>
</tbody>
</table>

### Second entry

<table>
<thead>
<tr>
<th>id</th>
<th>vlg_code</th>
<th>maize_harvest</th>
<th>hhh_age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>42</td>
<td>1600</td>
<td>55</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>1900</td>
<td>48</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>1800</td>
<td>22</td>
</tr>
</tbody>
</table>
Stata “how to”

```stata
* Set global directories
global dropbox  "C:\Users\Maria\Dropbox\"
global project  "$dropbox\FC Training\Thursday Presentations\double data entry example\"

** Load first entry data
use "$project/example_first_entry.dta"
describe

** Compare to second entry data
cfout using "$project/example_second_entry.dta", id(id)
```
Stata “how to”

- Stata output

```stata
. ** Compare to second entry data
. cfout using "$project/example_second_entry.dta", id(id)

<table>
<thead>
<tr>
<th>Total Discrepancies: 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Data Points Compared: 9</td>
</tr>
<tr>
<td>Percent Discrepancies: 22.222 percent</td>
</tr>
</tbody>
</table>

(output written to discrepancy report.csv)
```

.
Stata “how to”

- Cfout produces discrepancy report
  - double data entry example/discrepancy report.csv
    
    | id | Question        | Master | Using |
    |----|-----------------|--------|-------|
    | 1  | maize_harvest   | 1500   | 1600  |
    | 2  | vlg_code        | 3      | 8     |

- Make corrections based on hard copies
  - Double entry example/discrepancies corrected.csv
    
    | id | question          | correct_value |
    |----|-------------------|---------------|
    | 1  | maize_harvest     | 1500          |
    | 2  | vlg_code          | 8             |
Use “readreplace” to make corrections

- .csv has to have only 3 columns in this order
  - id, question name, correct value
- In this case 1 change is correct because only 1 error was in first entry

```stata
**: ** Make corrections to data
  use "$project/example_first_entry.dta", clear
  readreplace using "$project/discrepancies corrected.csv", id(id)
  Total changes made: 1

**: ** Save reconciled data
  save "$project/example_reconciled_data.dta", replace
  (note: file c:\users\Maria\Dropbox\FC Training\Thursday Presentations\double data entry example\example_reconciled_data.dta not found)
  file c:\users\Maria\Dropbox\FC Training\Thursday Presentations\double data entry example\example_reconciled_data.dta saved
```

end of do-file
Stata “how to”

- `cfout` – some useful options
  - `nopunct`: ignores differences in punctuation and capitalization
  - `nostring`: do not compare any string variables
  - `lower`: convert all string variables to lower case before comparing

- `cfout` assumes that the variable specified in the `id` option uniquely identifies observations in both datasets.

- `cfout` does not compare
  - variables that have a different string/numeric type in both datasets.
  - variables that are different in all observations.
- `cfby` is another useful command
- `cfby` compares variables from the dataset in memory to variables from the using dataset and displays the discrepancy rates by a common variable
  - useful if you are doing data entry and want to get discrepancy rates of data entry clerks
- `cfby [varlist] using filename, id(varname) [options]`
FC role

- Help train data entry operators
- Ongoing supervision of data entry process
- Check the data!!
  - Start early! Do not wait until data entry is finished!
- Check number of entered questionnaires for each cluster compared to survey log
Check data

- Check IDs – no duplicates, no odd values
- Confirm each cluster entered in full
- Compare entered data to sample frame to verify response rate
So you want to run a household survey...

Step 1: contract data collection agency
Step 2: design questionnaire
Step 3: pilot questionnaire
Step 4: draft survey manual
Step 5: train enumerators
Step 6: plan field work
Step 7: monitor data collection
Step 8: design data entry system
Step 9: supervise data entry
Step 10: data management and cleaning