Data entry tips

The cleaning of the first few datasets has resulted in some very bloated reports and some that looked reasonable. Many of the problems of the bloated reports could have been avoided. Here we offer some tips on data entry.

1. Missing Data

The technical guidelines are clear:

- **– 8 (minus eight)** is to be used to indicate that the question “does not apply” to the circumstances of the respondent(s).
- **– 9 (minus nine)** is to be used for the alternative “I don’t now” or ‘”The respondent doesn’t know”. Naturally, one should aim to minimize use of this response, but in some cases it’s unavoidable.

This means that there are very few cases where data would be missing (= nothing is entered). In fact, the only time missing data is acceptable is when the table has a leading question and 0 is entered in the leading question (more on this below). Moreover, there is a big difference between a blank field and zero (0) when it comes to data analysis.

Consider the example. It comes from the village survey. See the key below the table for the variable labels. The dots (.) represent gaps in the raw data.

```
  +-------------------------------------------------------------------
  | villcode | dem_yrvill | dem_hhd | dem_hhd10 | dem_in | dem_ethnic
  |----------+------------+---------+-----------+--------+------------
  |       30 |       1968 |       . |        90 |      5 |          3
  |----------+------------+---------+-----------+--------+------------
  |       31 |       1923 |      55 |        31 |      . |          3
  |----------+------------+---------+-----------+--------+------------
  |       34 |          . |      90 |        68 |     15 |          4
  |----------+------------+---------+-----------+--------+------------
Key
dem_yrvill:   year village established
dem_hhd:      how many households live currently in this village
dem_hhd10:    how many households lived in this village 10 years ago
dem_in:       how many persons living here now moved in last 10 years
```

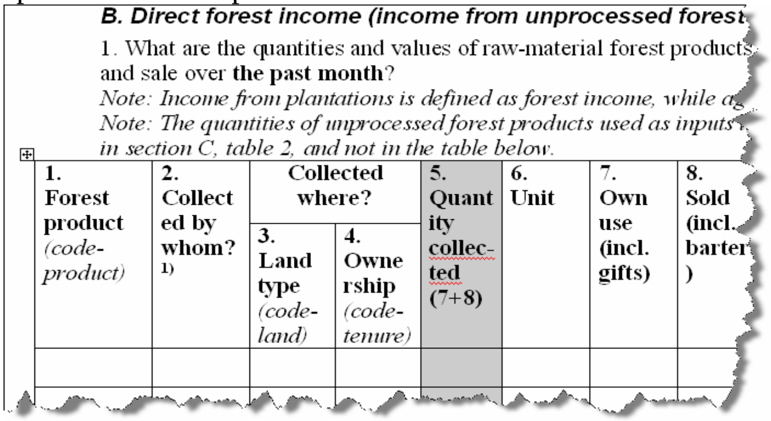
Why would these gaps be confusing? For **`dem_yrvill`**, it may be the case that, for village 34, it is not known (in which case, -9 should have been entered). It is also possible that, for village 30, the respondents did not know the number of households in the village (**`dem_hhd`**).

However, what does the gap for the **`dem_in`** (number of households that have moved in) for village 31 mean? Is it that zero households have moved or is this not known? The difference between 0 and missing (.) becomes very important when computing summary statistics as the later is excluded in such computations and therefore gives a more meaningful summary of the data.

Moral of the story: Enter as much data as you can and you will be rewarded with a small bug report.

2. Leading Questions:

The different modules ask products that households may or may not have. As an example, consider section B of the quarterly survey. This module asks about direct forest income from unprocessed forest products



B. Direct forest income (income from unprocessed forest

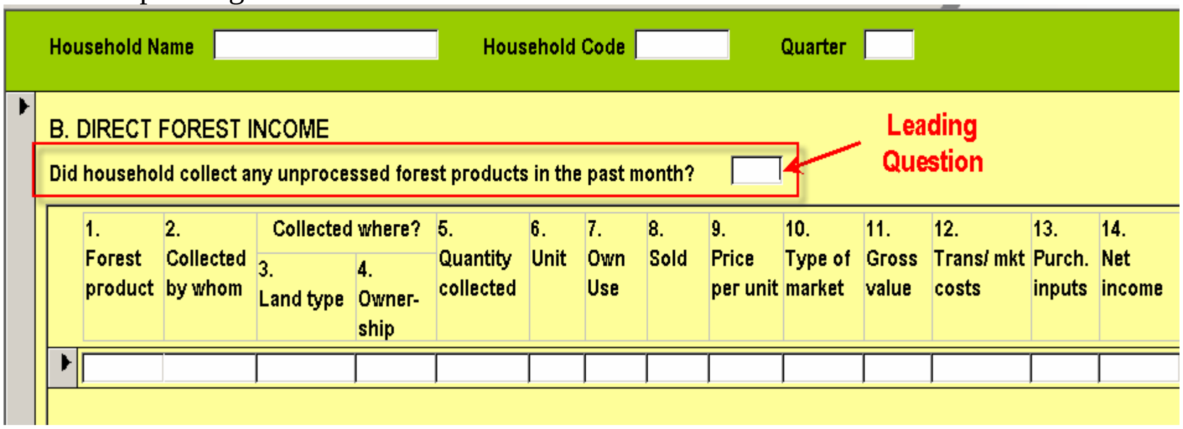
1. What are the quantities and values of raw-material forest products and sale over **the past month**?

Note: Income from plantations is defined as forest income, while a

Note: The quantities of unprocessed forest products used as inputs in section C, table 2, and not in the table below.

1. Forest product *(code-product)*	2. Collected by whom? 1)	Collected where?		5. Quantity collected (7+8)	6. Unit	7. Own use (incl. gifts)	8. Sold (incl. barter)
		3. Land type *(code-land)*	4. Ownership *(code-tenure)*				

The corresponding section in the database is



Household Name ▭ Household Code ▭ Quarter ▭

B. DIRECT FOREST INCOME

Did household collect any unprocessed forest products in the past month? ▭

1. Forest product	2. Collected by whom	Collected where?		5. Quantity collected	6. Unit	7. Own Use	8. Sold	9. Price per unit	10. Type of market	11. Gross value	12. Trans/ mkt costs	13. Purch. inputs	14. Net income
		3. Land type	4. Owner-ship										

Notice that, in the database, we added a leading question which asks, were any products collected (do you have any data on unprocessed forest products) and this is a YES NO (1, 0) question.

Part of the cleaning will check for inconsistencies, meaning that the data in the leading question should be consistent with the table that follows. If they had no products, then there should not be any data in the table that follows and vice versa. Here is an example of a bug report in which it was reported that products existed but no data was recorded

```
     Forest pdts in header but not included in table
+----------------------------------------------------+
|           househd | houscode | fup_lead | fup_pdt |
|------------------+----------+----------+---------|
| xxxxxxx xxxxxxxx |        4 |        1 |       . |
|------------------+----------+----------+---------|
|   xxxxx xxxxxxxx |       59 |        1 |       . |
|------------------+----------+----------+---------|
| xxxxxxx xxxxxxxx |       61 |        1 |       . |
|------------------+----------+----------+---------|
```

Fup_lead indicates that yes they had products (hence 1) but the products are missing.