

BISP Rapid Assessment Data: A Second Round of Data Quality Review

By Denis Nikitin
Consultant South Asia HD, WB

Overview:

IDS provided revised Rapid Assessment data sets on October 18, 2009. These revisions addressed problems with duplicate observations highlighted in an earlier note from October 4. The current dataset has mostly resolved the problem of duplicate observations and non-unique IDs. However, there are still some duplicate identifiers. The number of duplicate IDs is small and won't compromise the validity of inferences. Therefore, corrections of the *duplicate* data should only be done if it requires minimal effort.

DUPLICATES ISSUES

File Roster_j.dta

```
. duplicates tag sl_q_id sl_id_c, gen(repeat_id)
```

Duplicates in terms of sl_q_id sl_id_c

```
. tab repeat_id
```

repeat_id	Freq.	Percent	Cum.
0	13,011	99.98	99.98
1	2	0.02	100.00
Total	13,013	100.00	

```
. list sl_q_id sl_id_c if repeat_id==1
```

```
+-----+
| sl_q_id  sl_id_c |
+-----+
7586. |    1694      4 |
7587. |    1694      4 |
+-----+
```

File Main_i.dta

No duplication either in terms of all variables or at the level of HH ID + GENDER.

But there is one duplication in terms of HH ID + B_R_ID

```
. duplicates tag m_q_id b_r_id , gen(dup2)
```

Duplicates in terms of m_q_id b_r_id

```
.
end of do-file
```

```
. list m_q_id n_bene  r_bene b_r_id drop  gender age  n_resp if dup2==1
```

```
+-----+
| m_q_id      n_bene  r_bene  b_r_id  drop  gender  age      n_resp |
```

2515.	1562	MUNAWER JAN	Son	99	.	F	39	AZRA KHATOON
2516.	1562	MUNAWER JAN	Son	99	.	M	40	MUHAMMAD SAJJAD

File s5_b.dta

No duplication either in terms of all variables or HH + GENDER ID.

File s6_2431_b.dta

No duplication either in terms of all variables or HH + MONTH + GENDER ID.

File s6_3337_a.dta

No duplication either in terms of all variables or HH + GENDER ID.

File s7_1_3.g.dta

No duplication either in terms of all variables or HH + GENDER ID + ITEM.

File s78_4_12_e.dta

No duplication either in terms of all variables or HH + GENDER ID.

File exp_f_c.dta

Duplication in terms of HH + GENDER ID + ITEM, but can be corrected

```
. duplicates report q_id gender it_code
Duplicates in terms of q_id gender it_code
```

copies	observations	surplus
1	271705	0
2	2	1

```
. duplicates tag q_id gender it_code , gen(dup)
Duplicates in terms of q_id gender it_code
```

```
. list if dup==1
```

271510.	q_id	gender	it_code	none	p_q	p_v	w_q	w_v	o_q	o_v	r_q	r_v
	6025	Female	Salt (Iodise	None
	kpo_id				d_entry						dup	
	12				2882009						1	

271511.	q_id	gender	it_code	none	p_q	p_v	w_q	w_v	o_q	o_v	r_q	r_v
	6025	Female	Salt (Iodise	.	250	65
	kpo_id				d_entry						dup	
	12				2882009						1	

File exp_y_g.dta

No duplication either in terms of all variables or HH + GENDER + ITEM ID.