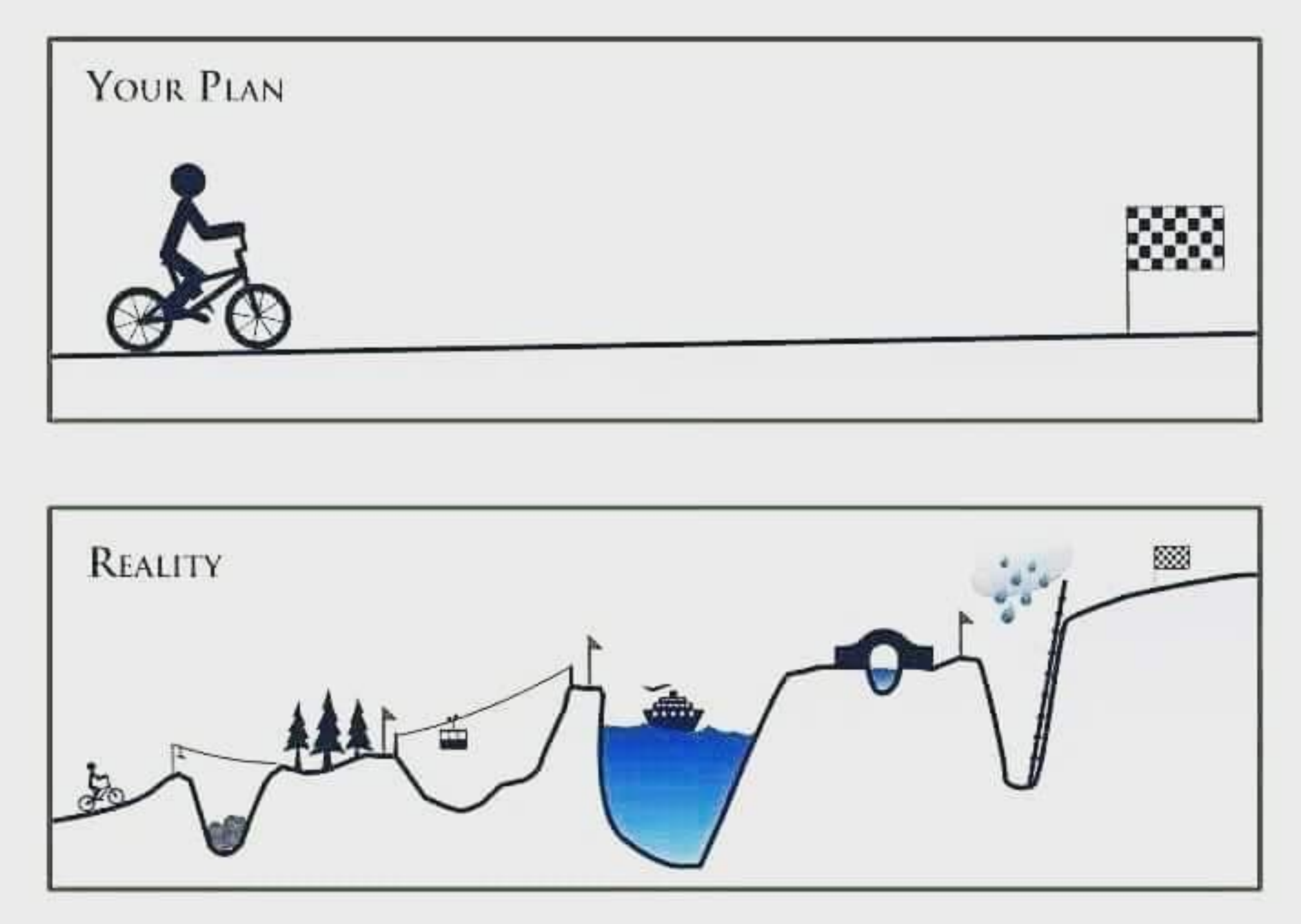
Issues in optical character recognition for statistical data capture

Arij Dekker
arijdekker@gmail.com

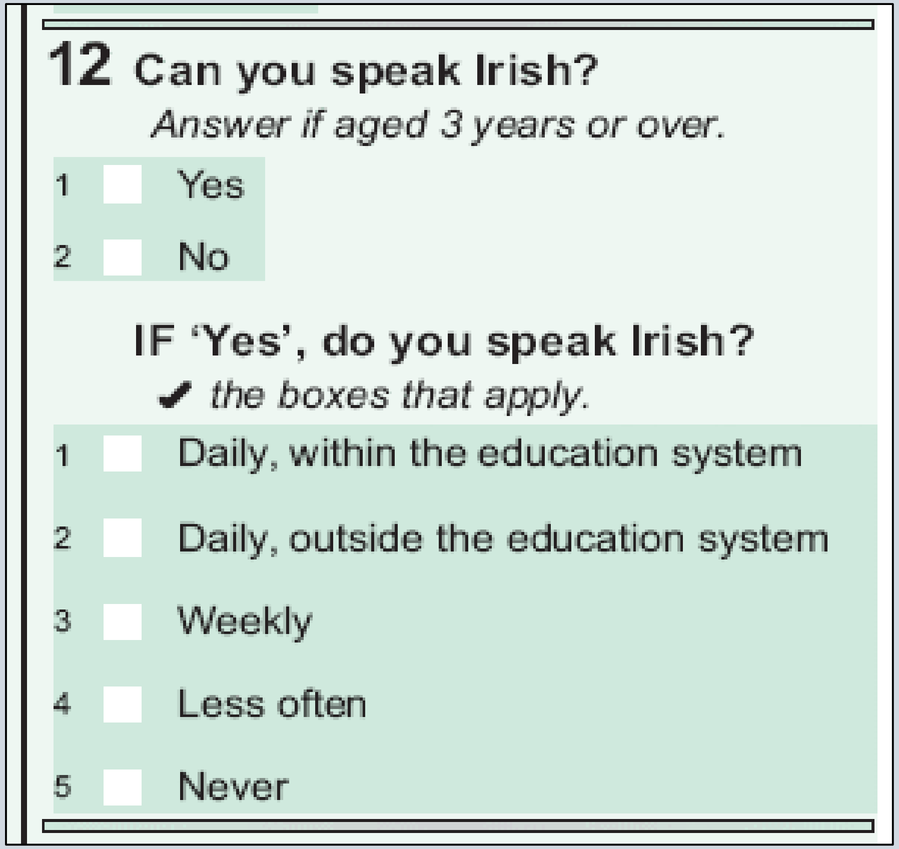
There will be problems



The choices

- Central keyboarding from paper
- Central scanning/reading from paper (OCR/ICR)
- Portable keyboard and store devices – no paper
- On-line enumeration or self-enumeration

OCR: Mark sensing



12 Can you speak Irish?

Answer if aged 3 years or over.

1 □ Yes

2 □ No

IF ‘Yes’, do you speak Irish?

✔ *the boxes that apply.*

1 □ Daily, within the education system

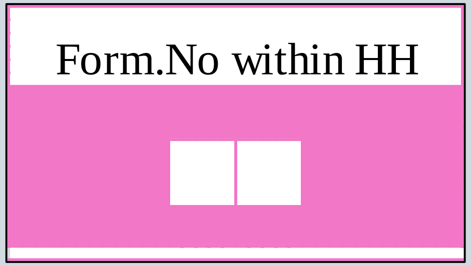
2 □ Daily, outside the education system

3 □ Weekly

4 □ Less often

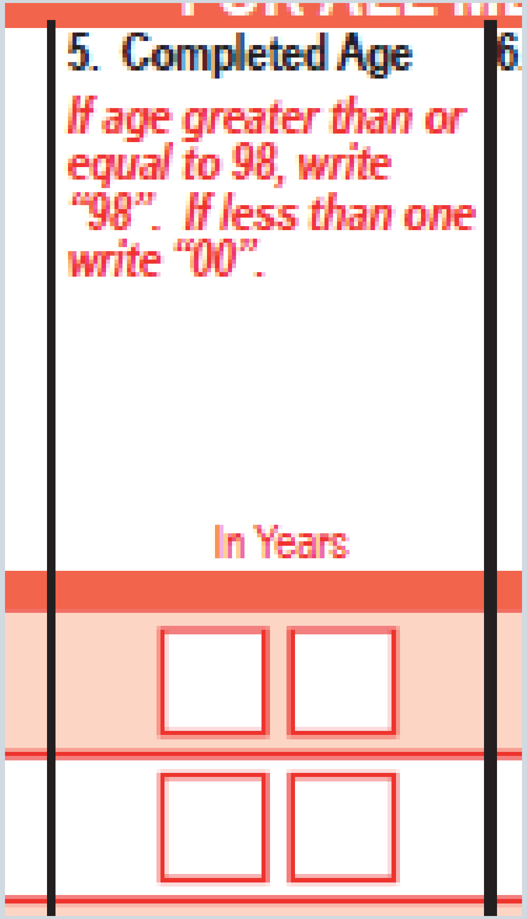
5 □ Never

OCR: Number recognition



Form.No within HH

□□



5. Completed Age

If age greater than or equal to 98, write “98”. If less than one write “00”.

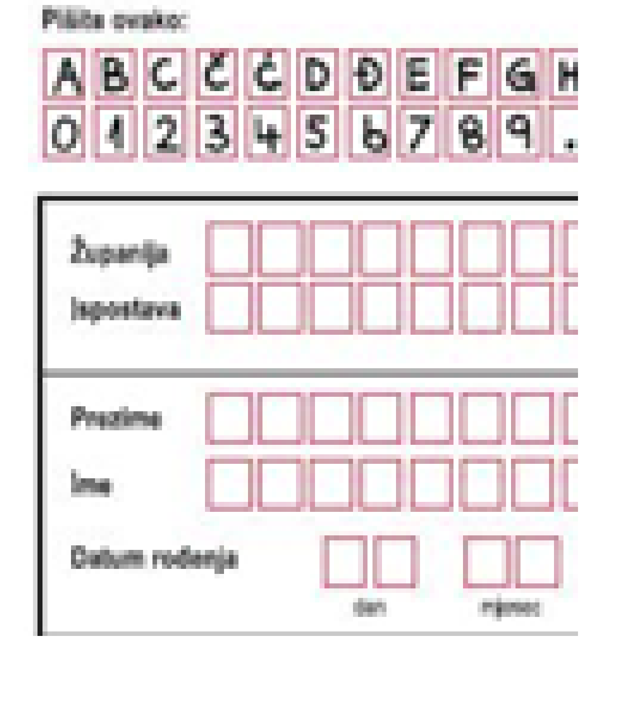
In Years

□□

□□

OCR: Working with text

Recognition



Pišite ovako:

A B C Č Ć D Đ E F G H

0 1 2 3 4 5 6 7 8 9 .

Županija

Ispostava

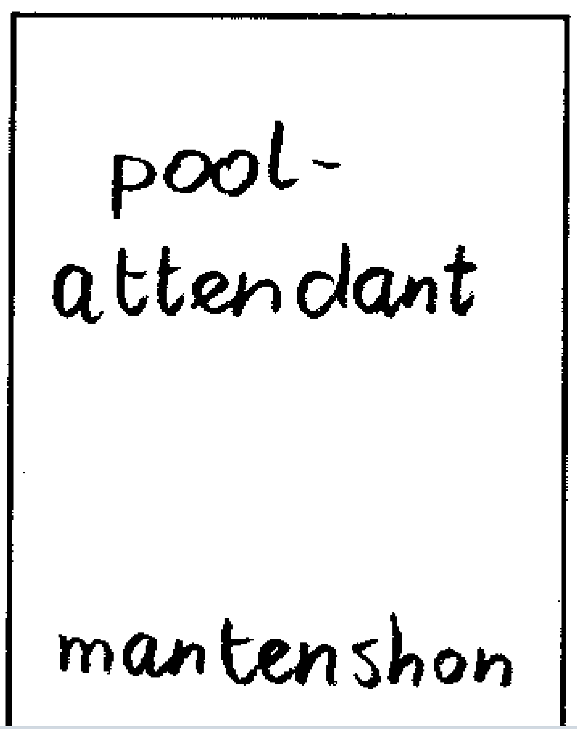
Prezime

Ime

Datum rođenja

dan mjesec

Computer Assisted Coding



pool-
attendant

mantenshon

One or several forms?

Single form:

- Cheaper
- Serves comfort
- Contractor choice

Dedicated forms:

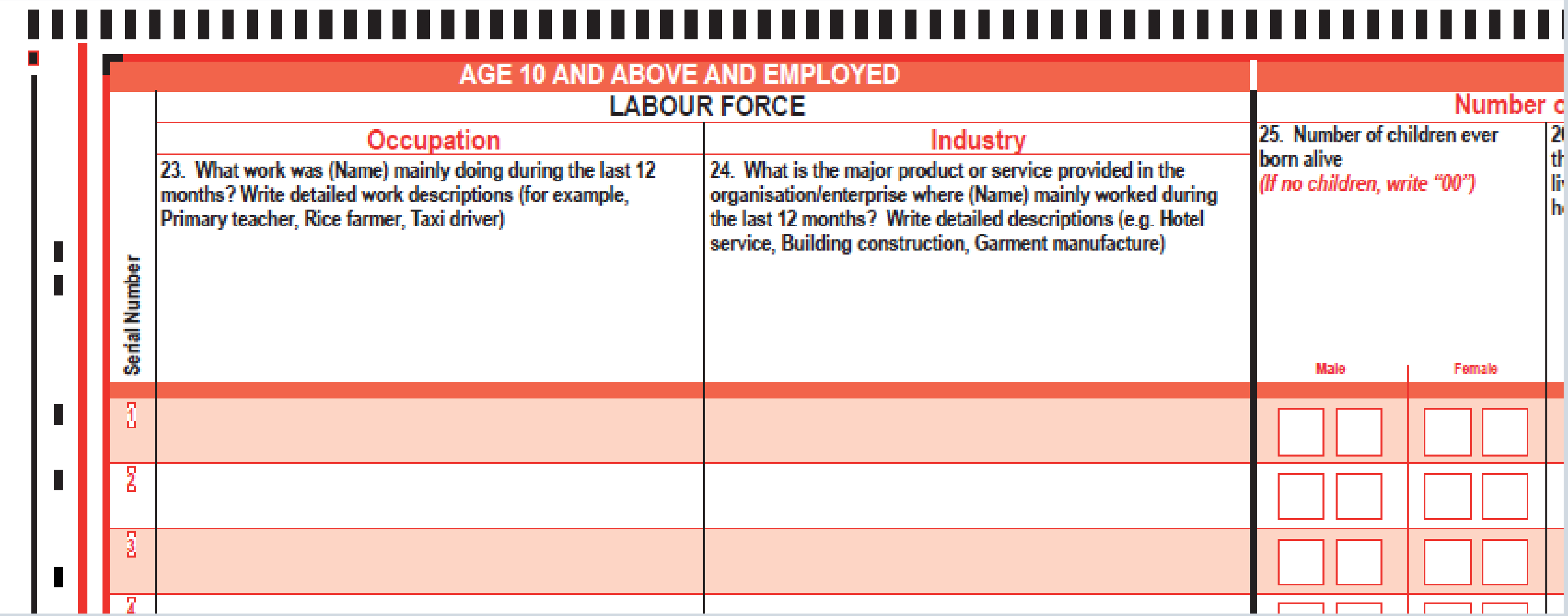
- Fewer “not applicables”
- Pre-numbering in first household sheet
- Improved control

OCR versus manual capture: issues

- Duplicate questionnaire Id's
- Range errors
- Inconsistencies, within form
- Inconsistencies between forms

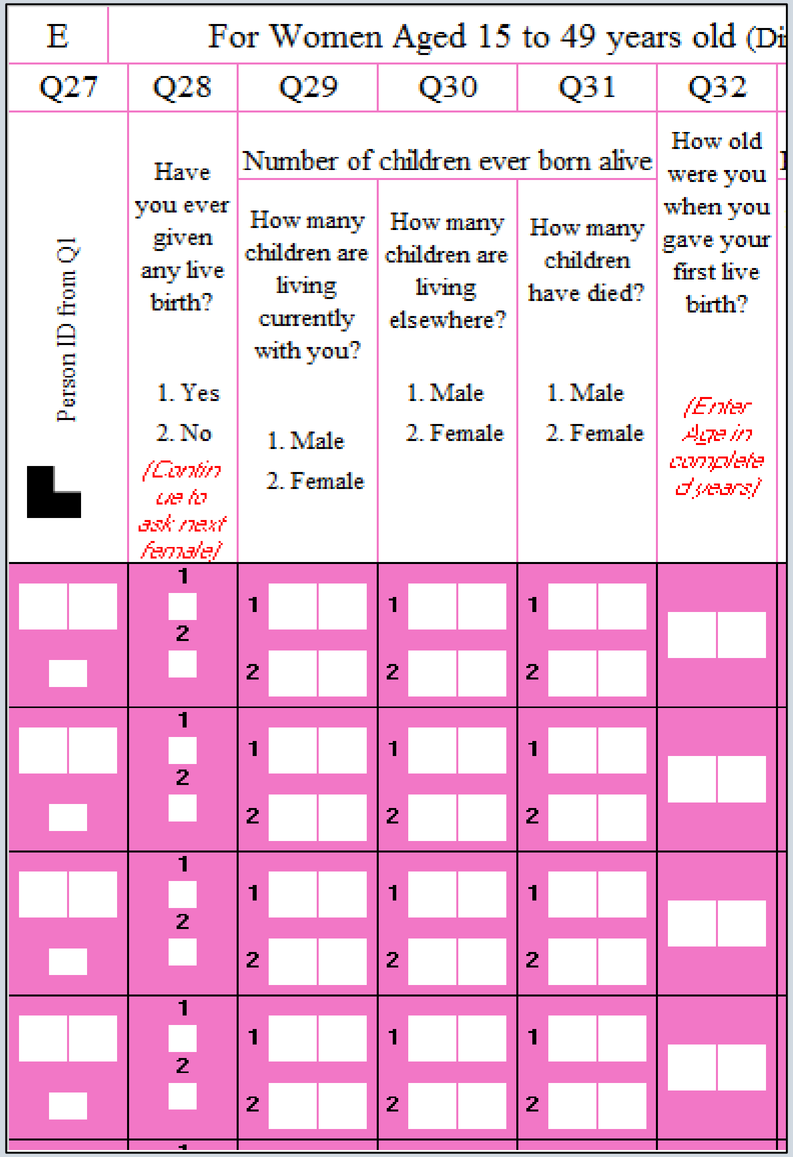
Reduced "hands-on" operator involvement during data capture places particular requirements on questionnaire design for OCR

Line or column diversion



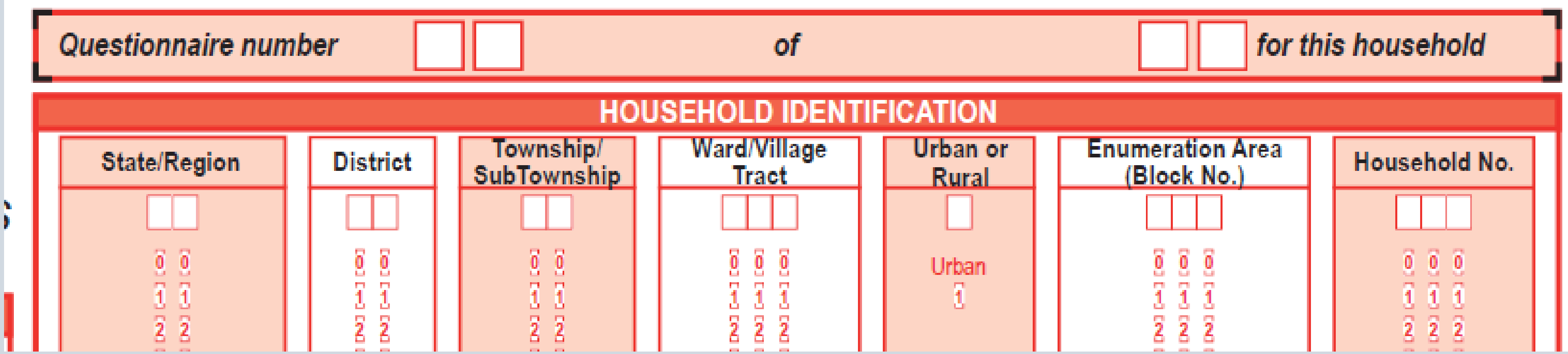
	AGE 10 AND ABOVE AND EMPLOYED LABOUR FORCE		Number c		
	Occupation	Industry			
Serial Number	23. What work was (Name) mainly doing during the last 12 months? Write detailed work descriptions (for example, Primary teacher, Rice farmer, Taxi driver)	24. What is the major product or service provided in the organisation/enterprise where (Name) mainly worked during the last 12 months? Write detailed descriptions (e.g. Hotel service, Building construction, Garment manufacture)	25. Number of children ever born alive *(If no children, write "00")* Male	Female	2 th li h
1			□□	□□	
2			□□	□□	
3			□□	□□	
4					

Repeating person id & Correcting mistakes



E	For Women Aged 15 to 49 years old (Di				
Q27	Q28	Q29	Q30	Q31	Q32
		Number of children ever born alive			
Person ID from Q1	Have you ever given any live birth? 1. Yes 2. No *(Continue to ask next female)*	How many children are living currently with you? 1. Male 2. Female	How many children are living elsewhere? 1. Male 2. Female	How many children have died? 1. Male 2. Female	How old were you when you gave your first live birth? *(Enter Age in completed years)*
□□ □	1 □ 2 □	1 □□ 2 □□	1 □□ 2 □□	1 □□ 2 □□	□□
□□ □	1 □ 2 □	1 □□ 2 □□	1 □□ 2 □□	1 □□ 2 □□	□□
□□ □	1 □ 2 □	1 □□ 2 □□	1 □□ 2 □□	1 □□ 2 □□	□□
□□ □	1 □ 2 □	1 □□ 2 □□	1 □□ 2 □□	1 □□ 2 □□	□□

Duplicate household id’s



Questionnaire number □□ *of* □□ *for this household*

HOUSEHOLD IDENTIFICATION

State/Region	District	Township/ SubTownship	Ward/Village Tract	Urban or Rural	Enumeration Area (Block No.)	Household No.
□□	□□	□□	□□□	□	□□□	□□□
0 0	0 0	0 0	0 0 0	Urban	0 0 0	0 0 0
1 1	1 1	1 1	1 1 1	1	1 1 1	1 1 1
2 2	2 2	2 2	2 2 2		2 2 2	2 2 2

Duplicate household id’s are the scourge of census files

Concluding remarks

- Involve IT staff in questionnaire design
- Test, test, test

- Prevent the risk of catastrophic failure