

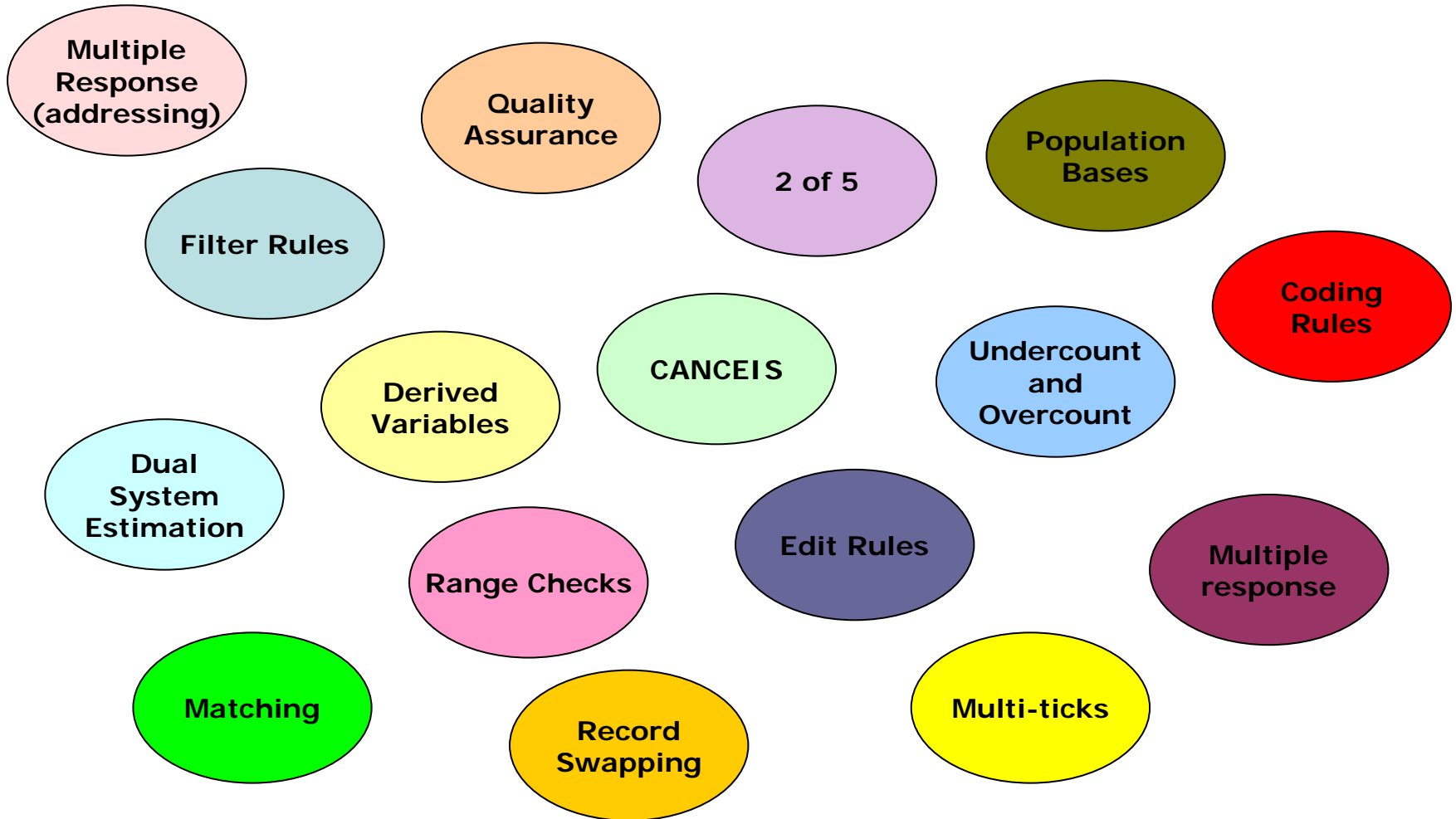
Data Processing

September 2011

Overview

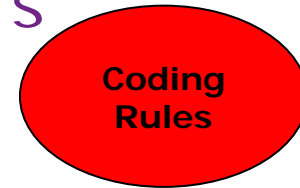
- What happens next?
 - An overview of processing
 - An overview of the timetable
- What does early analysis tell us about the quality of responses?

What happens next



Capture and Coding

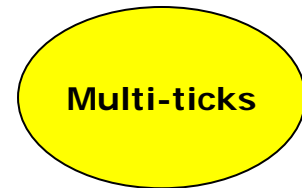
- Turning ticks and text into numbers
- What if people tick two boxes?
- Ticks AND text?



2 What is your sex?

Male Female

= 1 or 2



9 What is your country of birth?

England → Go to **13**

Wales → Go to **13**

Scotland → Go to **13**

Northern Ireland → Go to **13**

Republic of Ireland

Elsewhere, write in the current name of country

= 000 - 992

Data Cleaning (1)

- Removing False Persons

Problem identified in 2001

- Created by crossings out
- Created by dirty questionnaires/scanners



2 of 5

Remove using “2 of 5” rule

- Name (from individual questions or household members table) or Date of birth

AND

- One of: Name (from individual questions), Date of Birth, Sex, Marital Status, or Name (from household members table)

Data Cleaning (2)

- Multiple Responses from same household

Several types:

- Two or more questionnaires from same household

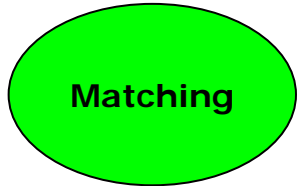
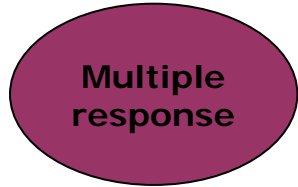
e.g.

- Two paper questionnaires returned
- Paper and Internet questionnaire returned

- Two or more responses for same person

e.g.

- Person on same questionnaire twice (or more!)
- Person on HH and Individual questionnaire
- Person on Paper and Internet questionnaire



Data Cleaning (3)

- Responses from different addresses but same household
 - Several types:
 - Different residential addresses for same household
e.g. Flat 1 and Flat A
 - Address on the CE and residential list
 - Identifying multiple responses:
 - Some identified during the field operation;
 - Further fieldwork on-going between July and Sept
 - Further analysis of address lists

**Multiple
Response
(addressing)**

Simple Processing

- Filter Rules
 - Identify incorrect / unnecessary responses
 - Prepare data for Edit and Imputation

7 Are you a schoolchild or student in full-time education?

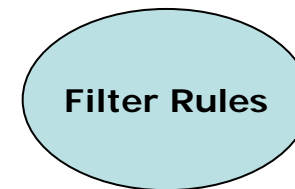
Yes No ➔ Goto **9**

8 During term time, do you live:

at the address on the front of this questionnaire?

at the address in question 5? ➔ Goto **43**

at another address? ➔ Goto **43**



Simple Processing

- First set of derived variables
 - Data required for processing

Derived Variables

3 What is your date of birth?

Day Month Year

<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
----------------------	----------------------	----------------------	----------------------	----------------------	----------------------

= Age

Edit and Imputation

- Resolves inconsistent responses
e.g. Someone is 5 years old and Married
- Imputes missing responses
 - Based on similar “donor” records
 - Households/people with similar characteristics
- Makes sure output database is complete

Derived Variables

Edit Rules

CANCEIS

Coverage Estimation and Adjustment

- Census Coverage Survey matched to census database
 - Automatic and clerical
- Used to estimate:
 - Missed households
 - Incomplete households
- Creates records for missing people
 - back through Edit and Imputation

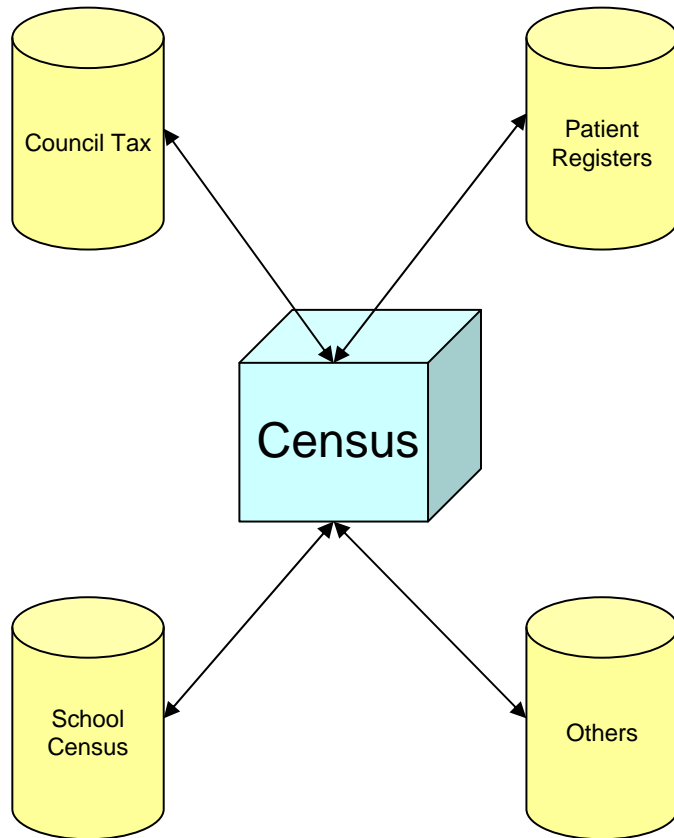
Undercount
and
Overcount

Matching

Dual
System
Estimation

(More on this later)

Quality Assurance



- High Level
 - Census compared to other data sources
 - Number of properties/people
 - People in specific groups
 - Not directly comparable
 - Definitional differences
 - Timeliness
- Detail
 - Does it look plausible?
 - Identify processing problems



(more on this later)

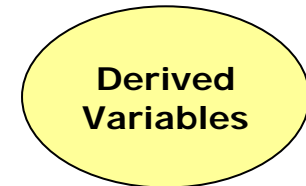
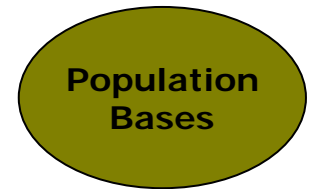
Disclosure Control

- Need to ensure outputs are confidential
 - Can't identify individuals
- Record swapping
 - Done successfully in 2001
 - Identify similar households and swap between areas
 - Mainly within LAs
 - Tables add up!
 - i.e. no post-tabular disclosure control



Outputs

- Main outputs for “usually resident” population
 - Everyone in UK for 12 months or more
- Some additional outputs on other bases
 - Daytime population
 - Out of term population



(More on this later)

Processing timeline

Introduction

Address
Register

Engagement
& Comms

Field
Operations

Online
Services

Data
Processing

Estimating the
Population

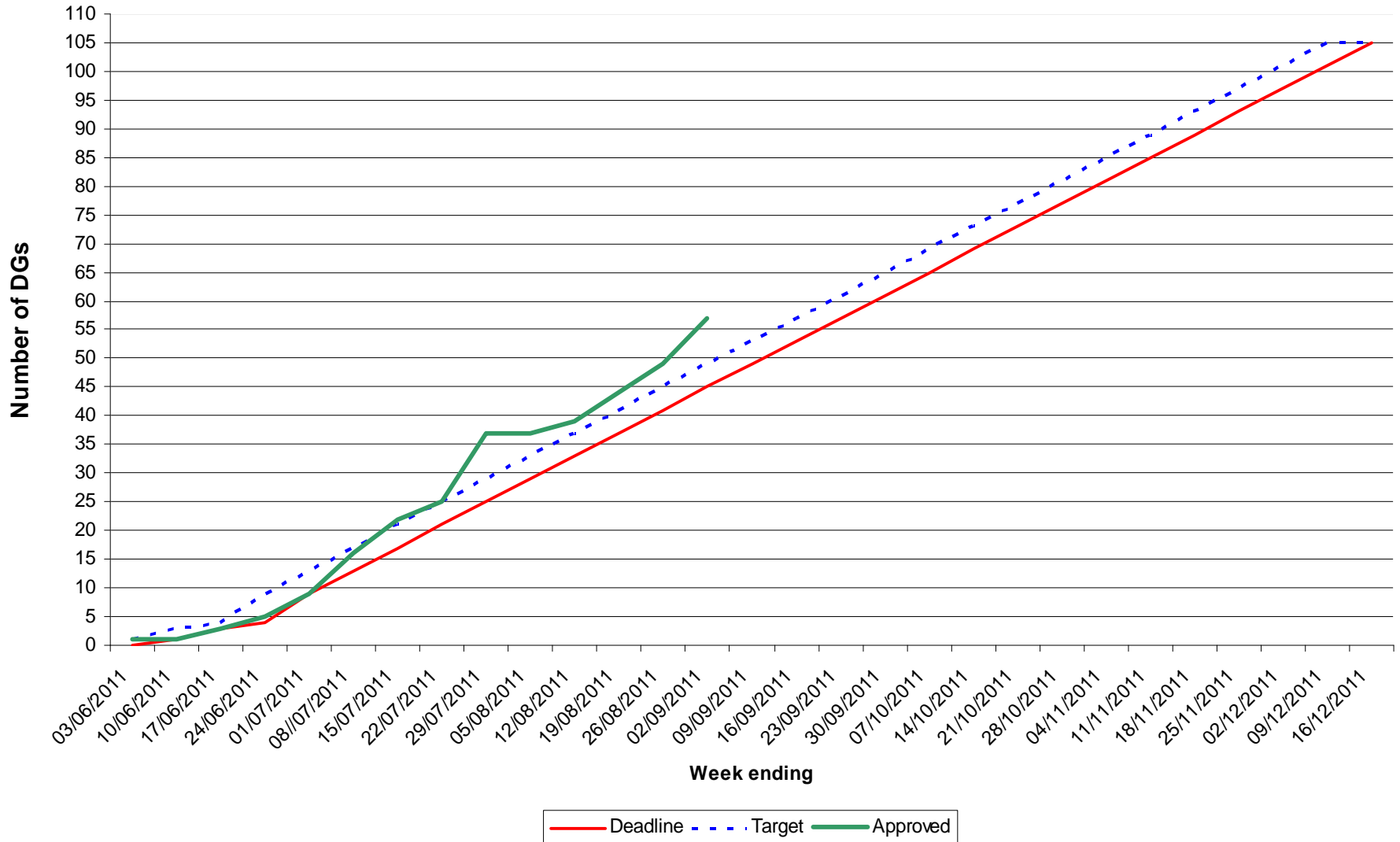
Census
Outputs

Summary

Timeline - key processing stages

Task Name	Q3 11		Q4 11			Q1 12			Q2 12			Q3 12			
	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	
Capture and coding (all DGs)	[Purple bar from Aug to Dec]														
Data load	[Light blue bar from Aug to Jan]														
RMR, filters & DVPs and edit & imputation	[Dark red bar from Aug to Jan]														
Coverage matching, estimation and adjustment	[Light green bar from Aug to Feb]														
Post coverage edit & imputation			[Blue bar from Sep to Mar]												
LA estimates signed off at QA Panel			[Orange bar from Sep to Apr]												
Output geographies finished									[Green bar from Apr to Aug]						
Statistical disclosure control									[Yellow bar from May to Sep]						
Public packs published												[Red bar from Jul to Sep]			

Capture and coding – delivery



What have we seen so far?

Introduction

Address
Register

Engagement
& Comms

Field
Operations

Online
Services

Data
Processing

Estimating the
Population

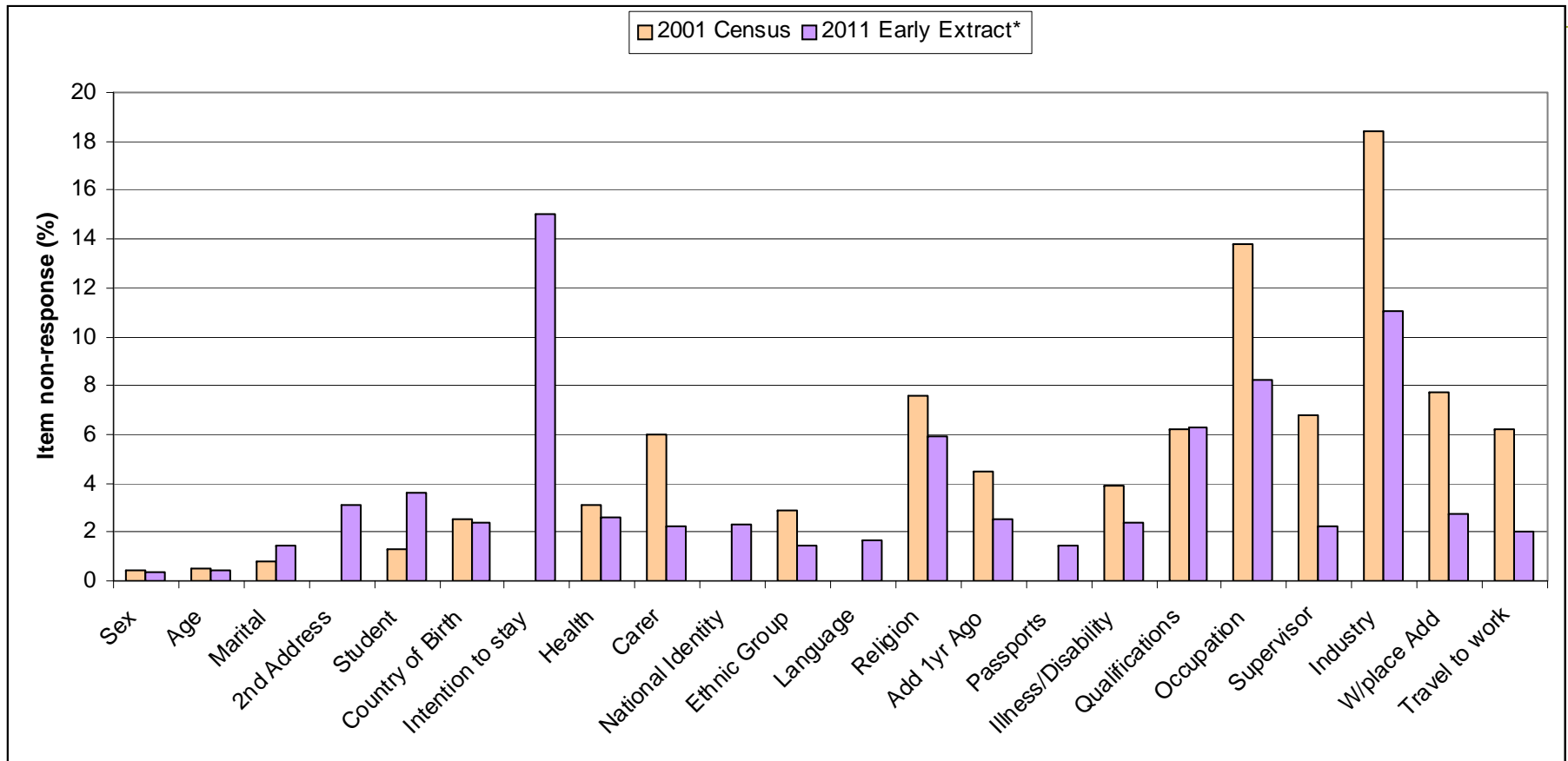
Census
Outputs

Summary

Capture and coding - Accuracy

Question	Target	Achieved (so far)
Sex	99.5%	99.96%
Year of Birth	99.95%	100%
Country of Birth	96%	99.86%
Religion	96%	99.09%
Ethnic Group	96%	98.96%
Language	96%	99.51%
Industry	88%	93.21%
Occupation	88%	93.87%
Workplace Address	85%	92.55%
Second Address	96%	98.8%

How well answered (so far)?



KPI 9: No question with greater than 19% non-response ✓

Obj 4: 70% of questions below 5% non-response ✓

* Provisional estimates of item non-response

Early assessment of progress and quality ...

- Some small problems identified during processing but relatively minor
 - Formatting difficulties
 - Some small coding inaccuracies
- Good quality respondent data
 - Item non-response low, new questions working well
- Processing and quality is an improvement over 2001 (So far!)