**Removing False Persons (RFP) process for the 2019 Census Rehearsal and 2021 England and Wales Census**

**Scott Redgwell – Census Processing**

**Introduction**

The following paper provides an analysis conducted on the Remove False Persons (RFP) process from the 2011 Census and 2017 Census Test, including an overview and evaluation of its requirements, analysis of the performance of the process, an overview of potential options and a recommended approach for use in the 2019 Census Rehearsal.

**Introduction to the RFP process**

Research carried out on 1999 Census rehearsal data recognised that individual records were being captured in error due to the following reasons:

* Tick recognition software picking up marks due to print quality & handling
* Dust on the scanners registered as ticks
* Crossings out on the page registering as ticks
* Writing 'NA' etc, into name field being recorded as a response
* Skipping pages and continuing answering on the section for a different person

These findings resulted in the introduction of the RFP process for the 2001 Census, which aimed to remove potential overcount prior to estimation. The RFP process specified that a person response needed to have answered 2 out of 4 specific variables to be retained and considered a genuine response, with any records failing this check being flagged for removal from the final data. The 4 variables were:

* Name (in individual section)
* Sex
* Data of Birth
* Marital Status

**2001 Census – RFP performance**

A review carried out on the 2001 Census RFP process showed that 3.3 million persons, or 6.3%, were removed as being false persons in England and Wales. It was believed that this process was removing too many genuine persons; this was confirmed by subsequent research conducted on records which failed this rule from 2001. This research reviewed the images of the scanned paper questionnaires to identify reasons for why the record was removed.

The results identified that the largest category of records that were removed were those who responded with their name elsewhere on the questionnaire but didn't enter it again on the person section of the form. This resulted in the specification being updated to the '2 of 5' rule for 2011 and required that, for a person record to be counted as a genuine response and kept in the data, the following information must have been present on the record:

* name (from individual questions or household members table) or Date of birth, and
* at least one other of: Name (from individual questions), Date of birth, Sex, Marital status, or Name (from household members table). Name in both person and household section did not count as passing.

In addition for 2011, improvements were made in the data capture procedures, meaning that fewer responses should have been captured by processing errors (poor printing quality, dust on the scanner etc). Also, the 2011 Census was the first to allow respondents to complete their return online, this functionality meant that non-response could be highlighted to the respondent as they progressed through the questionnaire, minimising records likely to fail RFP.

**2011 Census – RFP performance**

Because of the updates to the process, the RFP stage in 2011 removed fewer records, 982,000, or 1.8%, for England and Wales. As anticipated, most persons failing the ‘2 of 5’ rule responded via paper.

Table 1 shows the numbers of false persons, from paper responses, broken down by the total number of person questions completed. As expected, this table shows that most false persons completed little other data and can be assumed are being removed as intended. However, there is a rise in the proportion of false persons where between 15 and 19 person questions have been answered. It was identified that these data warrant further investigation to ensure the RFP rules are working as intended and not removing too many people.

*Table 1 - 2011 false persons (paper) by number of person questions answered*

|  |  |  |
| --- | --- | --- |
| Questions answered | Number of false persons | Percent |
| 0 to 4 | 548,785 | 66.14 |
| 5 to 9 | 84,780 | 10.22 |
| 10 to 14 | 47,616 | 5.74 |
| 15 to 19 | 100,356 | 12.09 |
| 20 to 24 | 45,497 | 5.48 |
| 25 and over | 2,751 | 0.33 |
| Total | 829,785 | 100 |

The remaining false person records from 2011 were submitted online. Most of these were from forced submitted questionnaires and 99% of all online false persons only answered 1 person question (the Name variable which has been automatically populated from the household section of the questionnaire). Due to the lack of extra data from online false persons, they were not considered for analysis.

**Results and discussion**

**Analysis 1 – 2017 Test images**

Due to 2011 Census images being unavailable, 2017 Test data were used for this piece of analysis. This was not considered to be a problem as the 2017 Test was thought to more representative of the likely environment for the 2021 Census, with a much greater proportion of respondents completing their questionnaires online.

The number of characters entered for every question of each person record were totalled and used to group the 2017 Test data into 4 categories. A tick-box would have been a single character whereas a write-in of ‘FRANCE’ would be six characters. The data were split into separate groups as follows:

* 9 or fewer characters
* between 10 and 14 characters
* between 15 and 19 characters
* 20 or more characters

 Table 2 shows the numbers of false persons from the 2017 Test, which were in each category.

*Table 2 - Total count of false persons for 2017 Test by category*

|  |  |  |
| --- | --- | --- |
| Category | Count | Percent |
| 1 | 1,358 | 71.44 |
| 2 | 50 | 2.63 |
| 3 | 58 | 3.05 |
| 4 | 435 | 22.88 |
| Total | 1,901 | 100.00 |

Responses from category 1 were not considered necessary for the analysis due to the small amount of data contained in each individual record. A random sample of 40 person IDs were taken from categories 2 to 4 and the scanned image of the questionnaire was manually viewed to ascertain the reason why this ID failed the RFP rule. These reasons were one of the following:

|  |  |  |
| --- | --- | --- |
| **Reason** | **Description** | **Desired outcome?** |
| **Duplicate person** | Responded multiple times for themselves | Yes |
| **Name in H3 only** | Respondent entered name on the household section, but only answered a limited number of other questions | Yes – in 2011, are they really false? |
| **No name on form** | Did not answer name in household or individual form, but answered other questions | Yes – in 2011, are they really false? |
| **Turned over pages** | Appears respondent turned more than one page on the questionnaire | Yes – in 2011, are they really false? |
| **Crossings out** | Respondent had crossed out a question or section and these marks had been captured as a response & scanning / capture issues | Yes |

**Categories 2 and 3**

The results from groups 2 and 3 showed very similar patterns. Table 3 shows a breakdown of the reasons for failing the RFP rules by the different sampled categories.

Table 3 - Reason for why record failed RFP by category (2017 Test – sampled records)

|  |  |
| --- | --- |
|  | **Reason for failing RFP (%)** |
| **Category** | **Duplicate person** | **Name in H3 only** | **No name on form** | **Turned over pages** | **Crossings out** | **Total** |
| **2** | 75 | 12.5 | 7.5 | 5 | 0 | 100 |
| **3** | 62.5 | 25 | 5 | 0 | 7.5 | 100 |
| **4** | 12.5 | 50 | 7.5 | 25 | 5 | 100 |

Most of the cases in category 2 (75%) and category 3 (62.5%) were identified as being duplicate persons. These are cases which should be removed from the data, but which would not be removed at any other processing stage, suggesting that for these categories RFP rules are working. Assessment of the questionnaires show that these cases of duplicates are usually person 2+ from a one-person household, where person 1 is real and often of pensionable age, and the remaining people are often partially completed with duplicate information from person one.

**Category 4**

Responses in category 4 are relatively well completed, unlike with categories 2 and 3, the false person is much more likely to be person 1 in the household (63.4%). 50% of cases sampled in category 4 provided a name in the household section of the form but didn’t answer 1 of the other questions required to pass RFP, suggesting they may have concerns regarding data security. A further 25% failed due to turning pages and missing the first page of a person section, which contains the important demographic questions required to pass the RFP rule.

**Analysis 2 – inclusion of ethnic group**

The next piece of analysis looked at extending the 2 of 5 rule by including ethnic group as an additional variable. Ethnic group was chosen as it is a key variable from an outputs perspective, it was answered in 70% of the cases who failed the 2011 Census RFP test on paper and it is on page two of the person section of the questionnaire, which would help to retain those who skipped pages.

Table 4 shows that, of those who failed RFP on paper in 2011, 17.8% would pass if ethnic group was included in the RFP rules. 14.7% would come from category 4, the category with the most complete responses.

Table 4 – Inclusion of ethnic group on those who failed RFP by category (2011 Census paper responses)



Table 5 shows that similar results are obtained when looking at 2017 Test data with 12.7% of those who failed RFP on paper passing with the inclusion of ethnic group.

Table 5 – Inclusion of ethnic group on those who failed RFP by category (2017 Test paper responses)



**Analysis 3 – 2011 Census detailed analysis**

Of the 829,785 people who responded on paper in 2011 but failed the RFP rule, 217,332 (26.2%) provided their name or date of birth and at least one other observed response. Table 6 shows that 74.29% of the 217,332 people completed at least 20% of the questionnaire, 57.01% responded to at least 40% of the questionnaire. This shows that although these respondents have failed the RFP test, they have answered a fair amount of the questionnaire suggesting they may be genuine responses.

Table 6 – Percentage of questionnaire completed by those who failed RFP but responded to either name or date of birth and at least one other question (2011 Census paper responses)



Table 7 shows the percentage of observed responses, by question, of the 217,332 who provided either their name or date of birth and at least one other observed response. As expected, questions from page one of the person section have a low completion rate, however questions from page 2 are well completed with most having an observed response for roughly 70% of the responses. Of all the person questions, national identity was answered the most at 76.7%. Page three of the questionnaire is also well completed with, as expected, the labour market set of questions that are placed across pages 3 and 4 having a lower completion rate due to a combination of these questions not requiring an answer from all respondents and the natural drop off that happens with survey completion.

Table 7 – Percentage of observed responses, by question, of those who failed RFP but responded to either name or date of birth and at least one other question (2011 Census paper responses)

**Analysis 4 – Implication on the Edit and Imputation (E&I) demography module**

Table 8 provides an indicator to show the questions used in the E&I demography module and their respective rate of completion by the 217,332 people who failed RFP but provided either their name or date of birth and at least one other observed response in the 2011 Census on paper.

One consideration when looking to expand the number of variables in the RFP rule is to look at the implications of using variables from the demography module as a potential 6th variable. The use of variables with a D\* indicator in table 8 could potentially result in the loss of that captured information later in processing, as this value could be punched out in later E&I modules to make newly imputed values consistent. Table 8 shows that disability is the only variable that could potentially result in the loss of a fair amount of real captured data. Therefore, the implications of using a variable from the demography module are considered to be minimal.

Table 8 – Percentage of observed responses, by question, of those who failed RFP but responded to either name or date of birth and at least one other question, including an indicator for inclusion in the Edit and Imputation demography module (2011 Census paper responses)





**Removing False Persons working group - meeting held 12/12/2018**

The above analyses were presented and discussed at the meeting, it was agreed that any decision regarding altering the rules for the RFP process is based on a trade-off between taking out responses from the census who are true (false positives) and leaving in responses who are false (false negatives).

It was agreed that although the analyses suggest the RFP rule may be taking out people who are true, it is widely believed that it is easier to account for undercoverage and much harder to account for overcoverage.

**Recommendation from the working group**

After consideration of the analysis, it was recommended that the 2021 RFP process continues to use the 2 of 5 rule from 2011 and the tool is built in a flexible way to allow for live analysis of the performance of the process so that any decision to alter the rule can be easily made and implemented during the operational phase.

In addition, a household level variable will be derived to record how many false persons have been removed from each household and a person level variable derived to contain a binary string showing completion of the questionnaire. This may be for the whole questionnaire or a subset of key questions.

**Next steps**

Development of the flexible RFP tool in the ONS Data Access Platform is under way and progressing well.

Census Processing are carrying out a final piece of RFP analysis looking at 2011 Census data to assess the consequence of making the RFP rule stricter, whereby to pass the rule both name and date of birth are required. Results shall be available at the end of February.

Methodology were actioned at the February CRAG meeting to consider whether the 2 of 5 rule is optimal for matching.