General Register Office for Scotland
information about Scotland's people

Statistical Evaluation of the 2006 Census Test in Scotland : Methodology Review

July 2008

Table of Contents

1. Introduction 3
2. 2006 Census Test Findings 4
3. Issues for Investigation 6
4. Methodology Review 7
5. Results of Analysis 12
6. Conclusion 23
Appendix A – Description of Analysis Variables 24
Appendix B – Frequency Tables 25
Appendix C - Logistic Regression Analyses 27
Appendix D - SAS programs 37
Appendix E - Glossary of terms 45

1. Introduction

1.1 The purpose of the 2006 Census Test was to test a range of strategies, procedures and instructions to inform future decisions on how to conduct the 2011 Census. The 2006 Census Test Evaluation, which was published in Spring 2007, included a statistical evaluation of the returned census forms.

1.2 The 2006 Census Test was run across five areas in Scotland covering about 50,000 households, which were purposefully chosen for the Test because each presented particular enumeration challenges. The areas were selected to cover urban, rural and semi-rural locations. Each area was also identified as containing hard to enumerate groups, including those with a high ethnic diversity, large numbers of asylum seekers, location within deprived areas, and those with low occupation rates such as crofts and holiday homes.

1.3 Two strategies for delivery of the Census Test forms to households were tested. Half of the households within each Enumeration District (ED) had forms hand delivered by an Enumerator, and the other half had the forms posted out to them. Forms were to be posted back or the completed forms could be collected by the Enumerator. These are referred to as the enumeration groups.

1.4 Similarly, half of the households within each ED received a form with a household income question included whilst the other half received a form without this question. These are referred to as the income groups.

1.5 The purpose of this review is to re-evaluate the survey strategy. In particular, we wish to evaluate the purposive sampling method used, and to assess the potential lack of randomness in both the choice of enumeration areas and the assignment of these areas to treatment groups (combination of enumeration group and income group). Whilst it is recognised that the results of the 2006 Census Test cannot be generalised to Scotland as a whole, it was accepted that the differential levels of undercount experienced in previous censuses may be repeated and so an assessment of suitable enumeration techniques to target hard-to-reach groups was required.

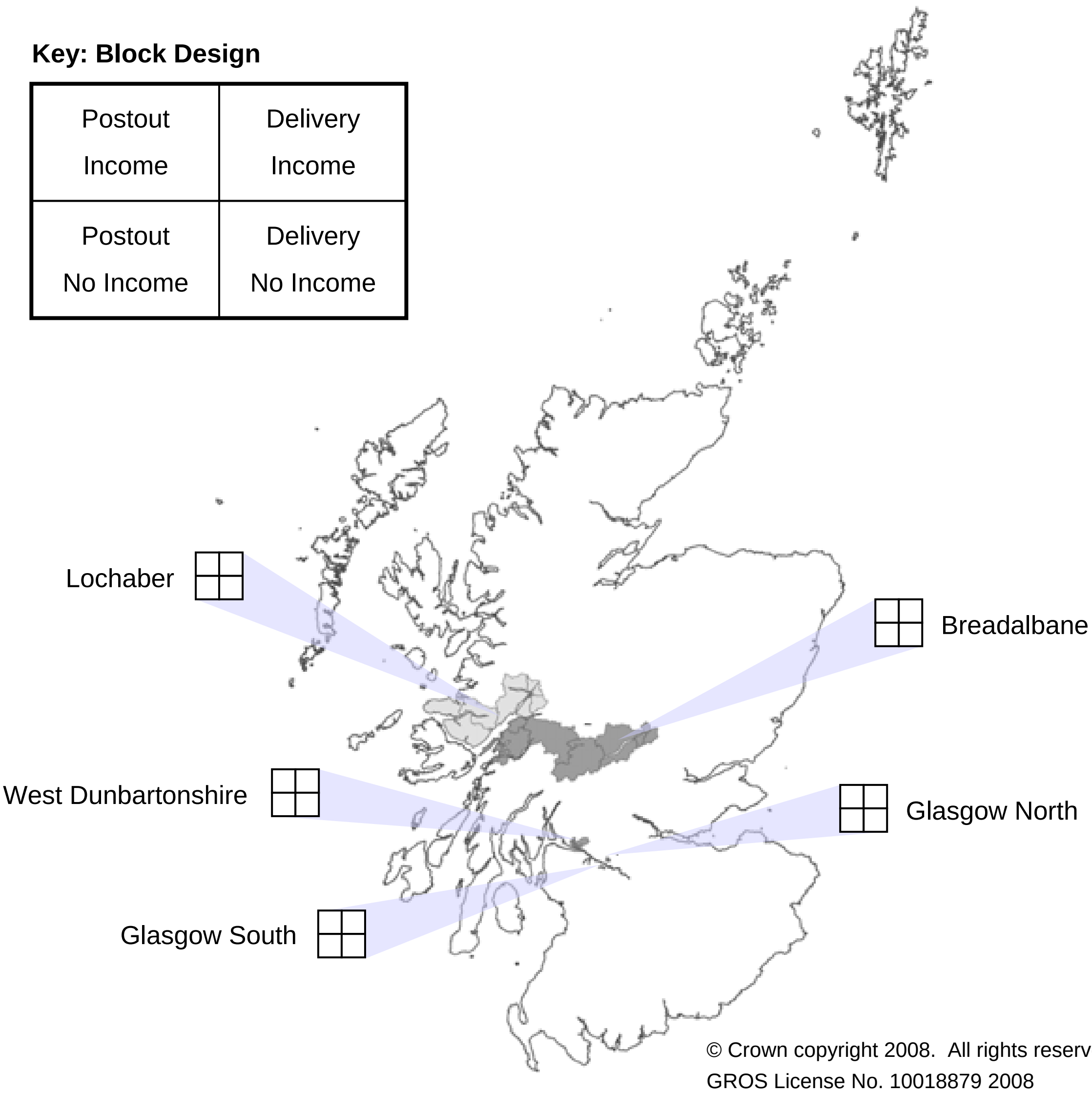
1.6 As such tests traditionally use a random sampling framework (choosing enumeration areas across Scotland at random), it was felt that such hard-to-reach groups were not likely to occur on a frequent enough basis in certain areas to allow the General Register Office for Scotland (GROS) to fully explore them. The 2001 Census demonstrated that differential undercount resulted from a wide variety of demographic, socio-economic and housing factors. Therefore, the Test sought to perform a qualitative analysis of households particularly prone to be undercounted with the view of allowing the implementation of schemes that could maximise coverage in the future. Primarily, the focus of the Test was twofold; to evaluate the impact on response of postal delivery (as opposed to hand delivery) and in the inclusion of an income question.

2. 2006 Census Test Findings

2.1 The Test was conducted in five areas in Scotland – in broad terms an area of deprivation, two multi-ethnic areas and two rural areas. The main differentiating aspect of the 2006 Census Test from other pre-census evaluative exercises was that in each of the chosen areas a complete census of households was attempted. The hope was that this would allow the collection of exploratory data from population subgroups that would not readily be found in a small-scale random survey.

2.2 The Census Test employed a two-factor design, treating the five enumeration areas as “blocks”. This allowed different combinations of enumeration strategy and income question inclusion to be tested across all of the areas. The first factor was the enumeration method and has two levels – postout/postback and delivery/postback – in both cases, the completed questionnaires were posted back but were either posted out or hand-delivered to the households. The second factor was income (whether the household income question was asked or not).

Figure 1: Block Design of the 2006 Census Test Design Variables



Key: Block Design

Postout Income	Delivery Income
Postout No Income	Delivery No Income

2.3 The two enumeration methods were trialled in order to investigate the link between delivery methods and census coverage and quality. Census-user consultations have increasingly shown that a question on income has been one of the most wanted questions absent from previous censuses. The supportive argument is that information on income is essential in determining the number of people who are socially excluded or deprived. Currently, estimates of these groups are determined using a number of proxy measures, which although fairly accurate, suffer from issues of reliability for small areas and subpopulations. On the other hand, the general public as a whole have been judged to be averse to compulsorily providing details of their income. Therefore, the income question was included in the Test to determine if respondents had issues about answering questions about their household income, and if any information collected would be of a sufficiently usable quality to inform government initiatives and to meet user needs.

2.4 The results of the Test from the original statistical evaluation report[1] showed that response rates for households receiving hand delivered questionnaires were higher – the response rate was 46% for postout and 53% for hand delivery. However, contrary to expectation, the households receiving a form with an income question had an apparent higher level of response than those without (48% compared to 44%). Even when the data is broken down by enumeration area, similar patterns emerge – for Glasgow North it is 34% compared to 29%, for Glasgow South it is 41% compared to 39%, for West Dunbartonshire it is 54% compared to 53%, for Lochaber it is 65% compared to 58% and finally the response for households with the income question in Breadalbane is 63% compared to 54% without the income question. However, subsequent analysis of the Test data identified two groups of responding households which affected more detailed statistical analysis – see section 4.1 for details.

2.5 The general outcomes of the Census Test were that the presence of an income question did not detrimentally affect the level of household response, and households that had their forms hand delivered by enumerators tended to have a higher response rate. It was decided that these preliminary results need to be further investigated given that the sampling scheme used to collect the data is somewhat complex. In order to make any conclusive inferences from the results, the sample design needs to be more fully incorporated into the analysis. The purpose of this methodology review, therefore, is to provide further insight into these results from this standpoint.

[1] http://www.gro-scotland.gov.uk/census/censushm2011/2006-census-test/2006-census-test-evaluation.html

3. Issues for Investigation

3.1 The ‘unusual’ results of the Scottish Census Test, in that some of the results run counter to expectations, merited further investigation. The results have also been the subject of some level of criticism from the Office for National Statistics, who are responsible for running the census in England and Wales. The main criticisms, which will be looked at in this methodology review, can be summarised into two points:

Issue 1: Lack of inference due to the non-randomness in the selection of areas

3.2 The five census areas in the 2006 Census Test were chosen because they were thought to contain people who were found to be particularly difficult to enumerate in the previous (2001) census. A similar non-random selection of test areas has been applied to the England and Wales Census Test; however, the difference is that in England and Wales the Test was not solely focused on extreme areas. Furthermore, the England and Wales Test was designed in such a way as to allow weighting of the results to allow inferences to be made to England and Wales as a whole.

Issue 2: Introduction of intra-Enumeration District correlations

3.3 It is possible that the sub-areas within the Enumeration Districts (ED) were too similar (especially given that entire neighbourhoods were assigned the same treatment effects, without randomisation of the households). Indeed, this failure to randomise the allocations may have had an adverse effect on the results.

3.4 It follows that any inferences about the difference (or lack thereof) between the treatment effects, even within the limited Test areas, are currently unsafe. There is the likelihood of intra-ED correlations (for example, clustering of sub-areas may mean responding or non-responding households have similar views which differ from others), therefore it is required to take this (neighbourhood) clustering into account in the analysis. Any failure to take proper account of this clustering in the design will most probably have led to overstating of the significance of the findings. Most of the statistical procedures used in the analysis of sample surveys assume, by default, that there is randomisation.

3.5 Both of the above points are concerned with the loss of efficiency due to the purposive nature of the Test’s design. In the 2006 Scottish Census Test sample design, similarities between households in each cluster will exist – more so when compared to a similar randomised sample of households.

4. Methodology Review

4.1 Number of Households

4.1.1 The data used for this analysis differs slightly from those used in the original evaluation. This is because two sets of households included in the original evaluation have now been excluded:

- Households that requested a replacement questionnaire. Only one type of questionnaire (with income question) was provided, and since such people have already gone to the trouble of asking for a replacement, they may be expected to be more likely to make a return – thus biasing the results.

- Households that were not on the original geography database (a defined list of households), but were found by enumerators. We know how many of these returned a questionnaire but not how many received one which they then did not return.

4.1.2 The revised data set consists of 51,101 households, compared to targeted delivery of 51,663 households (difference of 562) in the original evaluation. There were 562 households that requested and returned a replacement questionnaire, and for this analysis these have been removed from the delivery and response totals. There was a further 428 responses from households that were not on the original geography database. As such, the total number of received responses, for the purposes of this analysis, has been revised to 22,775, from 23,765 in the original evaluation (a difference of 990).

4.1.3 With these changes, the earlier results reported in section 2.4 were revised for the purposes of this analysis. Response rates for households receiving hand delivered questionnaires were higher – the response rate was 43% for postout and 46% for hand delivery. Also, the households receiving a form with an income question had a 44% response rate, while those who had no income question had a 45% response rate. There were differences in the latter comparison by enumeration area, with the group receiving a form with an income question having lower response rates in some areas but higher response rates in others.

4.2 Logistic Regression Analysis

4.2.1 In an information gathering exercise such as the Scottish Census Test, the objective is to investigate the effect of the different factors (enumeration district, enumeration method, income) on the response. The resulting data can be sorted out into distinct (non-overlapping) groups (or categories) on the basis of the response. When the response has two categories then this is referred to as a binary response.

4.2.2 For categorical data with binary responses, logistic regression modelling investigates the relationship between the binary response variable and the explanatory variables (those which may help explain why such a response was given). With reference to the Scottish Census Test, the logistic model is useful in predicting whether a household in the Test would, or would not, respond.

4.2.3 Logistic regression modelling is preferred to other regression modelling procedures because it makes no assumptions about the distribution of the explanatory variables. Therefore, it is of interest to model an event that happens (such as a household responding to the Test) with probability p, against a non-event which happens (such as a household not responding to the Test) with probability 1-p.

4.2.4 A ‘normal’ logistic analysis assumes that the data has been collected under randomisation, so that the explanatory variables explain most of the variation in the responses. However, this is not appropriate when data has been collected from complex survey designs which include stratification, clustering and unequal probability weighting. Failure to take account of the sampling design in such cases can lead to wrong estimates and inappropriate standard errors.

4.2.5 Therefore, the work carried out here looks at including the sample design in the analysis of the differences between households that did or did not respond in the 2006 Scottish Census Test. The logistic analysis compares and contrasts the ‘event’, that a household responded, against the ‘non-event’, that a household did not respond to the Test. It is hoped that performing a logistic analysis on the Census Test data, with a special consideration for the Test Design, will go some way to addressing the issues raised above.

4.2.6 By undertaking such analysis, we are using as much information as available – the non-responders, in effect, tell their own tale. Knowing what drives non-response in the 2006 Test is key in coming up with strategies that can be implemented in ensuring that coverage in the 2011 Census is as near-complete as possible.

4.3 Scottish Index of Multiple Deprivation

4.3.1 The 2006 Scottish Index of Multiple Deprivation (SIMD) provides a relative ranking of small areas across Scotland allowing the most deprived areas to be identified. There are 6,505 datazones across Scotland, with the datazone ranked 1 being the most deprived and the datazone ranked 6,505 being the least deprived.

4.3.2 Since deprivation is a fairly abstract concept, it is measured using some carefully chosen proxy indicators. There are seven domains of deprivation measured in the 2006 SIMD – income, employment, crime, education, health, housing and access to services.

4.3.3 The 2006 SIMD, like all the previous tools meant to target socially excluded areas, has been designed to identify pockets of areas with high levels of deprivation on multiple domains. It is important to note that the SIMD does not measure affluence, and so a small area with a high SIMD rank is not necessarily affluent, but simply less deprived. The belief that more affluent people will be less likely to fill in the Census Test form cannot be directly confirmed from any analysis – however, it is possible to investigate the statement that less deprived areas find the income question intrusive.

4.3.4 The 2006 SIMD was the second deprivation index to be produced by the Scottish Government (the first being in 2004) that provided deprivation data at a datazone-level. Datazones are groups of census output areas that were designed to have populations of between 500 and 1,000 household residents. They were also designed to nest within local authorities. They are deemed to be the most stable geographies when it comes to disseminating government statistics, and therefore allow changes over time to be properly analysed.

4.3.5 Previous SIMDs were based on wards, which were found to have ever-changing boundaries and as such did not facilitate easy comparison between different years. Unlike datazones, wards had populations of between 500 to 10,000 residents and although designed to contain households with similar social characteristics, it is visibly clear that information presented at a datazone geographical level is more useful.

4.3.6 The results of the SIMD are often presented by categorising the datazone ranks into smaller groups. For example, a quintile divides a ranked set of data into five equal groups whilst a decile divides a ranked set of data into ten equal groups. The least deprived decile and quintile refer respectively to the least deprived 650 (i.e. 10%) and 1,301 (i.e. 20%) datazones. The analysis in this report makes use of the SIMD deciles and quintiles.

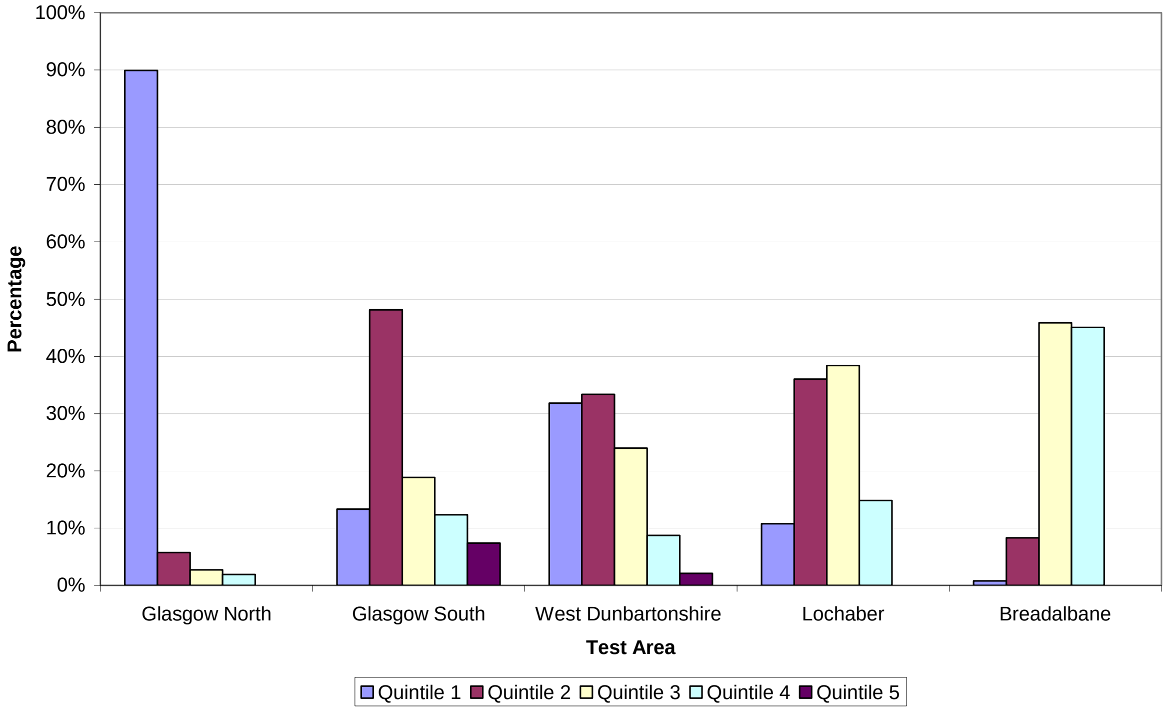
4.3.7 The deprivation data from the 2006 SIMD was linked to the Census Test data. For each household in the Test, it was possible to find the SIMD score and rank corresponding to the datazone in which the household was situated, and then determine the deprivation decile and quintile it belonged to. The intention was to find out if the distribution of households in each of the deprivation deciles/quintiles was sufficiently diverse in each enumeration district.

Table 1: Number of Households in each Census District by SIMD Quintiles

	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	Total
Glasgow North	10,633	677	323	224	0	11,857
Glasgow South	2,053	7,423	2,907	1,904	1,139	15,426
West Dunbartonshire	3,544	3,718	2,667	973	235	11,137
Lochaber	669	2,237	2,384	921	0	6,211
Breadalbane	51	537	2,966	2,916	0	6,470
Total	16,950	14,592	11,247	6,938	1,374	51,101

4.3.8 There is a lot of anecdotal evidence showing that harder-to-enumerate areas are more likely to be deprived (i.e. represented by being on the lower quintiles). Consequently, given that the enumeration districts were chosen because they contained populations that were difficult to enumerate in the 2001 Census (and a predominant number of these populations are economically depressed and socially excluded), it is expected that there will be some skew in the distribution. Table 1 and Figure 2 show that there is an uneven distribution (with some areas having more deprived areas than others), which is to be expected given the purposive nature of the Test.

Figure 2: Distribution of Households in each Census District by SIMD Quintiles



4.3.9 Of all the enumeration districts, only Glasgow South has a reasonably even distribution of households in each of the deprivation quintiles. Furthermore, one interesting point to notice is that the West Dunbartonshire test area, which was chosen because it had peripheral housing estates that had low response in 2001 Census, had households in all quintiles. This can be attributed to the way in which the Test was designed in that enumeration blocks were constructed using the estates as a starting point. As is often the case, pockets of deprived areas are located surrounded by more affluent areas. This meant that even though most of the households in West Dunbartonshire are fairly deprived, there are a number of more affluent households included (see Appendix A1 for the breakdown by deprivation deciles).

4.3.10 In order to accurately test the effect of the income question, it was required that half the households in the Test had the income question while the other half did not, taking into account the other design factors. The design of the Census Test did ensure that this held true for all the design factors, i.e. postout/delivery and enumeration district. On the other hand, by factoring in the deprivation quintiles a different picture emerges. Tables 2a and 2b show how income questions are distributed in each of the deprivation quintiles (Appendix A6 shows the distribution by deciles).

Table 2a: Number of Income/Non Income Questionnaires by SIMD Quintiles

		Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	Total
Income Asked	No	8,689	7,107	5,434	3,493	614	25,337
	Yes	8,261	7,485	5,813	3,445	760	25,764
	Total	16,950	14,592	11,247	6,938	1,374	51,101

Table 2b: Distribution of the Income/Non Income Questionnaires by SIMD Quintiles

		Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Income Asked	No	34.3%	28.0%	21.4%	13.8%	2.4%
	Yes	32.1%	29.1%	22.6%	13.4%	2.9%
	Difference	+2.2%	-1.1%	-1.2%	+0.4%	-0.5%

4.3.11 Hence, whilst the design made sure that the number of income/non-income questionnaires were roughly equally split (49.6% without income and 50.4% with income), when looking at the SIMD quintiles there is a definite difference in the proportion of forms sent out. 51% of households in Quintile 1 did not have the income question. For the other quintiles the percentage is 49% in Quintile 2, 48% in Quintile 3 and 50% in Quintile 4. The difference is much larger for the top quintile where 45% of households did not have the income question. This suggests that the SIMD should be included in the analyses as one of the design factors influencing the Census Test responses.

5. Results of Analysis

5.1 An exploration of the Test results was undertaken to answer the queries raised. All analysis was performed in SAS, a statistical analysis software package. A number of SAS programs were written for this purpose and are included in Appendix C. Logistic regression was used for the analysis, assessing whether or not a household returned their Census Test form (the response variable) against the design variables and other variables that were thought to have an influence on the response.

5.2 Two types of logistic regression were used: ‘normal’ logistic regression, which does not take the survey design into account but treats the data as if they came from a simple random sample, and a more sophisticated version (using the ‘surveylogistic’ procedure in SAS) which takes into account the effect of the clustered survey design. The objective of the model selection process is to build a logistic model that may be used to differentiate between households that responded and did not respond to the Census Test.

5.3 In all cases, the number of households (51,101) were divided into ‘event’ (22,775 households) and ‘non-event’ (28,326 households). The objective was to find out the defining characteristics of responding (events) and non-responding (non-events) households which can be used to fairly accurately differentiate between the households types.

5.4 Three different programs were written to perform different analysis (see Appendix C). For comparison purposes, program C1 implements the simplistic model (using the ‘logistic’ procedure in SAS) which uses the ‘raw’ counts (number of households) without any adjustment for the Census Test design. Program C2 (using the ‘surveylogistic’ procedure) stratifies the data by the SIMD quintiles, so here the assumption is that although the enumeration districts are fairly distinct, there are further effects that have not been taken into account – i.e. the deprivation of the areas. Finally, program C3 (again, using the ‘surveylogistic’ procedure) supposes that there is some clustering due to the enumeration districts being non-randomised, therefore there is the likelihood of correlation of responses between households within the same enumeration districts.

5.5 Analysis of the main effects

5.5.1 Tables 3 and 4 show that the presence of the income question does not significantly affect whether or not a household will respond. A p-value (indicated in the tables by the column marked ‘Pr > ChiSq’ and explained in Annex D) of 0.6387 can be interpreted to imply that the ‘income_asked’ variable will not have an effect on the propensity to respond of a household in 639 out of every 1,000 similar experiments. However, there is a strong effect for both the enumeration district (‘cen_dist’) and the enumeration method (‘postout’). For example, the enumeration method will not have an effect on the propensity to respond of a household in approximately 4 out of every 1,000 similar experiments.

Table 3: Analysis of Significance of Effects – With Income

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
cen_dist	4	601.1213	<.0001
postout	1	8.4684	0.0036
income_asked	1	0.2204	0.6387

Table 4: Analysis of Significance of Effects – With Income

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	0.1691	0.0541	9.7546	0.0018
cen_dist	**1**	1	-1.0458	0.0616	287.7726	<.0001
cen_dist	**2**	1	-0.6752	0.0578	136.5988	<.0001
cen_dist	**3**	1	-0.1341	0.0605	4.9068	0.0268
cen_dist	**4**	1	0.1587	0.0693	5.2365	0.0221
postout	**0**	1	0.1021	0.0351	8.4684	0.0036
income_asked	**0**	1	0.0165	0.0351	0.2204	0.6387

5.5.2 As the presence of the income question does not significantly affect whether or not a household will respond, it can be removed from the logistic model to allow for fuller investigation of the other effects. The logistic model, after removing ‘income_asked’ (see Appendix B2.4 for the SAS output), is given by

Model 1

$$\log(\frac{p}{1-p}) = 0.1772 - 1.0457 GlasgowN - 0.6752 GlasgowS - 0.1339 WestDunbarton + 0.1586 Lochaber + 0.1021 Postout$$

5.5.3 This model can be interpreted as the effect of each of the design variables (the census districts and the delivery method) and their individual effects on the propensity to respond.

5.5.4 The ‘intercept’ in Model 1 (represented by 0.1172) does not have any meaningful interpretation, apart from being the baseline from which comparison can be made – that is, the log-odds of returning the form for a household in Breadalbane (Census District 5), and keeping all the other design variables unchanged. The negative covariate estimates for Glasgow North, Glasgow South and West Dunbartonshire indicates that, in comparison to the baseline (i.e. Breadalbane), households in these areas were less likely to respond.

5.5.5 The odds ratios are much more conclusive (as shown in Table 5), when compared to the baseline. The table gives the odds ratio for Glasgow North (Census District 1) to be 0.351, which implies that a household in Breadalbane is roughly two thirds more likely to respond to the Test than a household in Glasgow North. Similarly, the estimated odds of a household in Glasgow South responding to the Test are half as high as for a household in Breadalbane. However, households in Lochaber were estimated to be much more likely to respond when compared to Breadalbane – they were 17% more likely (odds ratio point estimate of 1.172).

5.5.6 Controlling for all other design factors, delivery does seem to do marginally better than postout since it has an odds ratio of 1.107. This basically means that households with an enumerator delivered form were roughly 1.1 times more likely to respond than postal enumerated households. Indeed, this does follow what was found in the preliminary analysis.

Table 5: Odds Ratios of Model Estimates – Simple Logistic Model – Without Income

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
cen_dist 1 vs 5	0.351	0.313	0.395
cen_dist 2 vs 5	0.509	0.456	0.568
cen_dist 3 vs 5	0.875	0.779	0.982
cen_dist 4 vs 5	1.172	1.027	1.337
postout 0 vs 1	1.107	1.036	1.184

5.5.7 The effect of the income question has been excluded from this model, as the results in table 4 indicate that it is not statistically significant. So, although the number of forms returned with the income question were slightly more than those returned without, there is not enough evidence to discount that this result could have happened by chance.

5.5.8 Table 6 shows that there is no need to add an interaction term between the census district and enumeration method in the model (indicated by the ‘cen_dist*postout’ effect with a p-value of 0.7153). An interaction term may have indicated how the enumeration district and delivery method ‘interacted’ together to affect response.

Table 6: Investigation of an Interaction between Census District and Method – Without Income

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
cen_dist	4	302.4246	<.0001
postout	1	0.2704	0.6031
cen_dist*postout	4	2.1113	0.7153

5.5.9 Tables 7a and 7b show how well the model (without income) fits the data. The p-value of 0.2729 on the Hosmer-Lemeshow statistic is evidence that the model fits the data moderately well.

Table 7a: Goodness of Fit of Model

Partition for the Hosmer and Lemeshow Test					
Group	**Total**	**Event**		**Nonevent**	
		Observed	**Expected**	**Observed**	**Expected**
1	6,091	1,745	1,800.31	4,346	4,290.69
2	5,766	1,884	1,829.30	3,882	3,936.70
3	7,774	2,935	2,938.66	4,839	4,835.34
4	7,652	3,082	3,078.35	4,570	4,573.65
5	5,680	2,945	2,901.49	2,735	2,778.51
6	5.457	2,883	2,926.51	2,574	2,530.49
7	6,470	3,602	3,602.00	2,868	2,868.00
8	6,211	3,699	3,699.00	2,512	2,512.00

Table 7b: Goodness of Fit Statistic

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
7.5505	6	0.2729

5.6 Analysis by deprivation quintiles

5.6.1 The next modelling stage looked into controlling for the Scottish Index of Multiple Deprivation (SIMD) deprivation quintiles and the method used to create the enumeration districts (ED). It was expected, during the design, that there would not be much variation within the enumeration districts. However, due to the fact that treatments were assigned to households in close proximity to each other, the effect of clustering within treatment areas cannot be discounted.

5.6.2 The lack of randomisation of treatments was not expected to have much impact on the household response. The preliminary analysis did not adjust for this, and the size of the clusters suggest that it is possible that the sub-areal variation could have been inflated as a result. Consequently, three variables, ‘cluster_ED’, ‘cluster_SIMD’ and ‘cluster_ED_SIMD’, were created to investigate this. All three cluster variables were created using one of the SAS concatenate functions – basically joining the variable responses together.

5.6.3 The survey design was such that census districts were divided into enumeration districts, and each district was then divided into four so that the treatments (Income/Postout; No Income/Postout; Income/Delivery; No Income/Delivery) could be applied to the units. The purpose of ‘cluster_ED’ is to look at the effect of this clustering. In effect, knowing that a household was chosen to participate in the Test, and the treatment group it was in, meant that it was possible to ascertain that the next door neighbours were receiving the same treatments. It is this dependence that needs to be adjusted for in the analysis.

5.6.4 ‘cluster_ED’ was formed by joining the ‘unique_ED’, ‘income_asked’ and ‘postout’ variables. For example, the concatenated value ‘31611’ represents the cluster of households in unique_ED 316, with income (1) and with postout (1). Similarly, ‘31600’ represents the cluster of households in unique_ED 316, without income (0), and with delivery (0).

5.6.5 The variable ‘cluster_SIMD’ works in the same way, but this time it groups by the SIMD quintiles instead of the EDs. There are a total of 110 EDs – 24 in Glasgow North, 31 in Glasgow South, 23 in West Dunbartonshire, 16 in Lochaber and 16 in Breadalbane. The EDs were chosen to minimise the workloads of the enumerators and there were different criteria for each of the different census districts. This means that ED1 in Breadalbane is not comparable to, say, ED1 in Glasgow North. However, a household in SIMD quintile 1 in Breadalbane is more or less comparable to a household in the same quintile in Glasgow North, Glasgow South, West Dunbartonshire or Lochaber.

5.6.6 ‘cluster_SIMD’ was created by concatenating ‘income_asked’, ‘postout’ and the SIMD quintile. For example, the value ‘115’ represents the cluster of households with the income question asked (1), with postout (1) and in the least deprived quintile (i.e. simqui 5).

5.6.7 The variable ‘cluster_ED_SIMD’ was formed by concatenating ‘unique_ED’, ‘income_asked’, ‘postout’ and the SIMD variables. For example, the value '316115’ represents the cluster of households in unique_ED 316, with income (1), postout (1) and in the least deprived quintile (5).

5.6.8 Not surprisingly, the model with ‘cluster_SIMD’ is preferred since it has a more meaningful interpretation than ‘cluster_ED’, given that the SIMD quintiles across areas can be compared with ease. As the numbering on the EDs did not have any meaningful interpretation, the purpose of ‘cluster_ED_SIMD’ was to try and isolate the enumerator effects from the deprivation effects. However, it is still difficult to interpret, so ‘cluster_SIMD’ is the preferred clustering variable.

5.6.9 When the clusters are included in the model, the effect this has is to reduce the significance of the covariates. However, it cannot be fully discounted that the reduction in significance could be caused by the fact that the analysis is being carried out on smaller cell counts; for example the model clustering by the SIMD quintiles has 20 clusters, while there are 440 clusters when clustering by ED. Working with smaller counts could lead to larger standard errors and wider confidence intervals.

5.6.10 Tables 8 to 10 give the odds estimates of the model using ‘proc surveylogistic’ with an account taken of the stratified sampling employed using the SIMD quintiles.

Table 8: Odds Ratios[2] of Model Estimates – Stratified by SIMD Quintiles - No Cluster Effects

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
cen_dist 1 vs 5	0.351	0.330	0.374
cen_dist 2 vs 5	0.509	0.480	0.540
cen_dist 3 vs 5	0.875	0.822	0.930
cen_dist 4 vs 5	1.172	1.092	1.257
postout 0 vs 1	1.107	1.069	1.148

[2] Any odds ratio confidence interval that overlaps 1 (for example, going from 0.924 to 1.136) is representative of the fact that there is not enough evidence to reject the hypothesis of no effect.

Table 9: Odds Ratios of Model Estimates – Stratified by SIMD Quintiles – ED-cluster Model

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
cen_dist 1 vs 5	0.351	0.271	0.456
cen_dist 2 vs 5	0.509	0.425	0.610
cen_dist 3 vs 5	0.875	0.717	1.067
cen_dist 4 vs 5	1.172	0.948	1.449
postout 0 vs 1	1.107	1.019	1.204

Table 10: Odds Ratios of Model Estimates – Stratified by SIMD Quintiles – SIMD-cluster Model

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
cen_dist 1 vs 5	0.351	0.316	0.391
cen_dist 2 vs 5	0.509	0.449	0.577
cen_dist 3 vs 5	0.875	0.750	1.020
cen_dist 4 vs 5	1.172	0.992	1.384
postout 0 vs 1	1.107	0.991	1.238

5.6.11 Tables 8 to 10 demonstrate that the confidence intervals for the odds ratio estimates are narrower when stratified by the deprivation quintiles alone, as compared to the wider confidence intervals when accounting for clustering by ED or deprivation quintiles. However, the estimates of the odds ratios are the same which indicates that the effects on response for ‘income_asked’, ‘cen_dist’ and ‘postout’ are the same.

5.6.12 Table 10 shows that the odds ratio estimate confidence interval for the ‘postout’ variable overlaps 1, and this is suggestive of there being not enough evidence against the null hypothesis of no effect due to enumeration method. The p-value for the ‘postout’ parameter coefficient under this model is 0.0727 (see Table 11), which is not significant at the 5% level but is significant at the 10% level.

Table 11: Analysis of Significance of Effects – Stratified by SIMD Quintiles – SIMD-cluster Model

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-0.1110	0.0253	19.3052	<.0001
cen_dist	**1**	1	-0.7065	0.0299	559.2384	<.0001
cen_dist	**2**	1	-0.3360	0.0371	82.1134	<.0001
cen_dist	**3**	1	0.2053	0.0521	15.5251	<.0001
cen_dist	**4**	1	0.4978	0.0562	78.3999	<.0001
postout	**0**	1	0.0510	0.0284	3.2209	0.0727

5.6.13 Removing the ‘postout’ covariate from the model just leaves the census districts in the model shown in Table 12. The choice is now between the model with just the census districts, and the model with both the enumeration method and census districts.

Table 12: Odds Ratios of Model Estimates – SIMD-cluster Model (with only census district in the model)

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
cen_dist 1 vs 5	0.351	0.310	0.398
cen_dist 2 vs 5	0.509	0.449	0.578
cen_dist 3 vs 5	0.874	0.748	1.021
cen_dist 4 vs 5	1.172	0.992	1.386

5.6.14 In model selection, it is better to treat variables that are found to be significant at the 10% level, but not at the 5% level, with due care. It is sometimes possible that in the presence of other factors the variable becomes significant. This is certainly the case with the ‘postout’ variable. When the SIMD quintiles are included in the model, ‘postout’ is found to be significant (see Table 13).

Table 13: Type 3 Analysis of Effects for Model

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
cen_dist	4	254.2933	<.0001
simdqui	4	241.9161	<.0001
postout	1	18.2930	<.0001

5.6.15 The model with the census districts, enumeration method and SIMD quintiles is intuitive. There is firstly a difference due to the census districts because of the way in which households were chosen to participate in the Census Test. Secondly the enumeration method does have an effect on whether or not a household will respond. Finally, there is still some variation in the responses that can be attributed to the deprivation of the areas.

Table 14: Model including the SIMD Quintiles (simdqui) as a Covariate

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
cen_dist 1 vs 5	0.515	0.447	0.593
cen_dist 2 vs 5	0.557	0.505	0.614
cen_dist 3 vs 5	1.044	0.942	1.157
cen_dist 4 vs 5	1.299	1.135	1.487
postout 0 vs 1	1.101	1.054	1.151
simdqui 1 vs 5	0.484	0.429	0.546
simdqui 2 vs 5	0.653	0.586	0.727
simdqui 3 vs 5	0.685	0.614	0.765
simdqui 4 vs 5	0.819	0.738	0.908

5.6.16 As a result, owing to what was earlier mentioned about the difference between forms sent to households, when broken down by SIMD quintiles (see Table 2), the model that includes the SIMD as a covariate is preferred (see Appendix B3, Analysis of Maximum Likelihood Estimates).

Model 2:

$$\log(\frac{p}{1-p}) = -0.0017 - 0.4750 GlasgowN - 0.3967 GlasgowS + 0.2321 WDunbarton + 0.4507 Lochaber + 0.0482 Postout$$

5.6.17 Note that since the covariates are categorical with no distinct ordering (apart from the SIMD quintiles which are based on ranked datazones) a baseline needs to chosen to be used for a comparison. This was chosen to be households in the least deprived SIMD quintile in Breadalbane, holding all other factors unchanged. This model also takes into account the clustering that exists due to the sample design, using 'cluster_SIMD'.

5.6.18 By re-arranging this model, we are able to calculate the predictive probabilities. These are useful in showing what the chosen model predicts the response probability will be, given some household characteristics. The predictive probability of a household responding, p, can be found by re-arranging Model 2 and is given by

$$p = \frac{\exp(-0.0017 - 0.4750 GlasgowN - 0.3967 GlasgowS + 0.2321 WDunbarton + 0.4507 Lochaber + 0.0482 Postout)}{1 + \exp(-0.0017 - 0.4750 GlasgowN - 0.3967 GlasgowS + 0.2321 WDunbarton + 0.4507 Lochaber + 0.0482 Postout)}$$

5.6.19 The predictive probabilities can, more often than not, be more intuitive than the odds ratios, which are the default provided by SAS. The odds ratios, which basically present the ratio of the odds of an event occurring in one group to the odds of it occurring in another group, only allow comparison to the baseline, but the predictive probabilities can allow multiple comparisons.

Table 15: Predicted Probabilities under the Preferred Model with the SIMD Quintiles

	Quintile 1		Quintile 2		Quintile 3		Quintile 4		Quintile 5	
	With Income	**Without Income**	**With Income**	**Without Income**	**With Income**	**Without Income**	**With Income**	**Without Income**	**With Income**	**Without Income**
Breadalbane										
Postout	0.406	0.406	0.479	0.479	0.492	0.492	0.536	0.536	0.500	0.500
Delivery	0.418	0.418	0.491	0.491	0.504	0.504	0.548	0.548	0.512	0.512
Lochaber										
Postout	0.517	0.517	0.591	0.591	0.603	0.603	0.645	0.645	0.610	0.610
Delivery	0.529	0.529	0.603	0.603	0.614	0.614	0.656	0.656	0.622	0.622
West Dunbartonshire										
Postout	0.463	0.463	0.537	0.537	0.549	0.549	0.593	0.593	0.557	0.557
Delivery	0.475	0.475	0.549	0.549	0.561	0.561	0.605	0.605	0.569	0.569
Glasgow South										
Postout	0.315	0.315	0.382	0.382	0.394	0.394	0.437	0.437	0.402	0.402
Delivery	0.325	0.325	0.394	0.394	0.406	0.406	0.449	0.449	0.413	0.413
Glasgow North										
Postout	0.298	0.298	0.364	0.364	0.376	0.376	0.418	0.418	0.383	0.383
Delivery	0.308	0.308	0.375	0.375	0.387	0.387	0.430	0.430	0.394	0.394

5.6.20 The predictive probabilities are given in Table 15 above. Also, notice that the predicted probabilities for households which receive income and non-income questionnaires are the same since under the preferred model the presence of the income question does not have a significant influence on the household's propensity to respond.

5.6.21 The analysis has shown that there is evidence of clustering in the sample. Failure to fully account for this clustering can overstate the significance of the results because the associated standard errors may be inaccurate. There is also the likelihood of a lack of randomisation due to the geographic deprivation, that can be corrected for by including the SIMD deprivation measure directly in the model. Nevertheless, the conclusions – with and without accounting for the effect of clustering – are practically the same.

6. Conclusion

6.1 The 2006 Census Test was run across five areas in Scotland covering about 50,000 households, which were purposefully chosen for the Test because each presented particular enumeration challenges. The purposive nature of the Scottish 2006 Census Test means that it cannot readily be used to make inferences about the whole of Scotland, unlike say a fully randomised study.

6.2 However, the analysis undertaken for this report demonstrates that there is conclusive evidence that the income question did not have a significant effect on the household response. Further, the enumeration methodology did have a significant effect on whether or not a household responded. Households that had their forms enumerator-delivered had a greater propensity to respond than households whose forms were posted. Despite there being some clustering effects due to deprivation, and the way in which the enumeration districts were created, the results from the original analysis still remain pretty much unchanged.

6.3 The main issues investigated here lies in the fact that there was a lack of randomness in the Test design. The test design covered the assignment of the treatment effects (inclusion of income question, delivery method, and enumeration area) to the subjects (households). There is an increasing body of literature debating the merits and demerits of random assignment. It is true that in order to make any valid inferences to a larger population (than the sample covered) it is essential that an experiment is suitably randomised. Nevertheless, a dogged focus on randomising can detract from the main aims of an experiment, especially in cases where the populations under study have characteristics that make them of interest, and will not occur frequently enough under a randomised experiment. In such an experiment, even if it may not be possible to easily generalise the results to the whole of the population, inferences may be drawn about the characteristics that may be related to the outcome under study.

Appendix A – Description of Analysis Variables

The following are the variable definitions used in the analysis of this report.

Cen_dist	Census District	1	Glasgow North
		2	Glasgow South
		3	West Dunbartonshire
		4	Lochaber
		5	Breadalbane
Response	Was a response to the questionnaire received	0	No
		1	Yes
Income_asked	Was the income question included in the questionnaire	0	No
		1	Yes
Postout	Method of delivery of questionnaire	0	Postout
		1	Hand delivery

Appendix B – Frequency Tables

B1: Distribution of Census Districts by SIMD Deciles

Table of cen_dist by simddec											
cen_dist	**simddec**										**Total**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
1	6,914	3,719	677	0	323	0	224	0	0	0	11,857
2	754	1,299	3,745	3,678	1,345	1,562	1,089	815	951	188	15,426
3	937	2,607	1,751	1,967	901	1,766	736	237	235	0	11,137
4	1	668	977	1,260	822	1,562	566	355	0	0	6,211
5	51	0	0	537	747	2,219	1,470	1,446	0	0	6,470
Total	8,657	8,293	7,150	7,442	4,138	7,109	4,085	2,853	1,186	188	51,101

B2: Distribution of Census Districts by SIMD Quintiles

Table of cen_dist by simdqui						
cen_dist	**simdqui**					**Total**
	1	**2**	**3**	**4**	**5**	
1	10,633	677	323	224	0	11,857
2	2,053	7,423	2,907	1,904	1,139	15,426
3	3,544	3,718	2,667	973	235	11,137
4	669	2,237	2,384	921	0	6,211
5	51	537	2,966	2,916	0	6,470
Total	16,950	14,592	11,247	6,938	1,374	51,101

B3: Distribution of the Response Profile by SIMD Deciles

Table of response by simddec											
response	**simddec**										**Total**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
0	5,941	5,190	3,839	4,061	2,054	3,434	1,904	1,234	582	87	28,326
1	2,716	3,103	3,311	3,381	2,084	3,675	2,181	1,619	604	101	22,775
Total	8,657	8,293	7,150	7,442	4,138	7,109	4,085	2,853	1,186	188	51,101

B4: Distribution of the Response Profile by SIMD Quintiles

Table of response by simdqui						
response	simdqui					Total
	1	2	3	4	5	
0	11,131	7,900	5,488	3,138	669	28,326
1	5,819	6,692	5,759	3,800	705	22,775
Total	16,950	14,592	11,247	6,938	1,374	51,101

B5: Distribution of the Income/Non Income Questionnaires by SIMD Quintiles

Table of income_asked by simdqui						
income_asked	simdqui					Total
	1	2	3	4	5	
0	8,689	7,107	5,434	3,493	614	25,337
1	8,261	7,485	5,813	3,445	760	25,764
Total	16,950	14,592	11,247	6,938	1,374	51,101

B6: Distribution of the Income/Non Income Questionnaires by SIMD Deciles

Table of income_asked by simddec											
income_asked	simddec										Total
	1	2	3	4	5	6	7	8	9	10	
0	4,500	4,189	3,390	3,717	1,889	3,545	2,233	1,260	521	93	25,337
1	4,157	4,104	3,760	3,725	2,249	3,564	1,852	1,593	665	95	25,764
Total	8,657	8,293	7,150	7,442	4,138	7,109	4,085	2,853	1,186	188	51,101

Appendix C - Logistic Regression Analyses

C1: Data for Logistic Analysis (without any stratification or clustering)

Obs	cen_dist	postout	income_asked	r	n
1	1	0	0	925	2,880
2	1	0	1	959	2,886
3	1	1	0	853	3,011
4	1	1	1	892	3,080
5	2	0	0	1,579	3,797
6	2	0	1	1,503	3,855
7	2	1	0	1,442	3,828
8	2	1	1	1,493	3,946
9	3	0	0	1,489	2,744
10	3	0	1	1,394	2,713
11	3	1	0	1,537	2,856
12	3	1	1	1,408	2,824
13	4	0	0	846	1,461
14	4	0	1	1,061	1,669
15	4	1	0	924	1,562
16	4	1	1	868	1,519
17	5	0	0	894	1,581
18	5	0	1	919	1,638
19	5	1	0	848	1,617
20	5	1	1	941	1,634

r = number of forms received and n = number of forms delivered

C2 - The Logistic Procedure for the Unclustered (and Unstratified) Data

Notes useful in understanding SAS output [3]

a. The **Type 3 Analysis of Effects** makes sure that the effect of the variable does not depend on the order in which the variable is specified in the model. This becomes important when considering models with more than one parameter.

b. The **95% Wald CIs** provide a range in which the "true" parameter may lie, and if the interval includes 1 then there is not enough evidence to reject the null hypothesis of no difference (effectively that the particular regression coefficient is zero).

c. For each model parameter, it is required to test the null hypothesis of no effect, against an alternative. The Wald Chi-Squared test statistic is given and shows the calculated test statistic corresponding to the hypothesis that the given parameter's estimate is equal to zero.

d. **Pr>ChiSq** is the probability that the Wald Chi-Squared test statistic would be observed under the null hypothesis that a particular predictor's regression coefficient is zero. Therefore, for a given significance level Pr>ChiSq determines whether or not the null hypothesis can be rejected – if Pr>ChiSq is less than 0.05 then the null hypothesis can be rejected and the parameter estimate is considered statistically significant at the 5% level.

[3] http://www.ats.ucla.edu/STAT/sas/output/sas_ologit_output.htm

C2.1 Univariate Model – census district

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
cen_dist	4	416.1087	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.2279	0.0580	15.4492	<.0001
cen_dist	1	1	-1.0462	0.0741	199.3068	<.0001
cen_dist	2	1	-0.6749	0.0695	94.4476	<.0001
cen_dist	3	1	-0.1346	0.0728	3.4232	0.0643
cen_dist	4	1	0.1591	0.0834	3.6434	0.0563

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
cen_dist 1 vs 5	0.351	0.304	0.406
cen_dist 2 vs 5	0.509	0.444	0.583
cen_dist 3 vs 5	0.874	0.758	1.008
cen_dist 4 vs 5	1.172	0.996	1.381

C2.2 Univariate Model – postout

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
postout	1	0.2613	0.6092

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-0.2693	0.1428	3.5542	0.0594
postout	**0**	1	0.1036	0.2027	0.2613	0.6092

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
postout 0 vs 1	1.109	0.746	1.650

C2.3 Univariate Model – income

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
income_asked	1	0.0048	0.9447

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-0.2251	0.1437	2.4548	0.1172
income_asked	**0**	1	0.0141	0.2040	0.0048	0.9447

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
income_asked 0 vs 1	1.014	0.680	1.513

C2.4 Best Model – census district and postout

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
cen_dist	4	636.4877	<.0001
postout	1	8.9568	0.0028

Analysis of Maximum Likelihood Estimates						
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	0.1772	0.0498	12.6600	0.0004
cen_dist	**1**	1	-1.0457	0.0599	304.6939	<.0001
cen_dist	**2**	1	-0.6752	0.0561	144.6414	<.0001
cen_dist	**3**	1	-0.1339	0.0588	5.1850	0.0228
cen_dist	**4**	1	0.1586	0.0674	5.5364	0.0186
postout	**0**	1	0.1021	0.0341	8.9568	0.0028

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
cen_dist 1 vs 5	0.351	0.313	0.395
cen_dist 2 vs 5	0.509	0.456	0.568
cen_dist 3 vs 5	0.875	0.779	0.982
cen_dist 4 vs 5	1.172	1.027	1.337
postout 0 vs 1	1.107	1.036	1.184

C3: Output from the best model

Notes useful in understanding the SAS output [4]

a. **Percent Concordant -** A pair of observations with different observed responses is said to be concordant if the observation with the lower ordered response value has a lower predicted mean score than the observation with the higher ordered response value.

b. **Percent Discordant -** If the observation with the lower ordered response value has a higher predicted mean score than the observation with the higher ordered response value, then the pair is discordant

c. **Percent Tied -** If a pair of observations with different responses is neither concordant nor discordant, it is a tie.

d. **Pairs -** This is the total number of distinct pairs.

e. **Somer's D** - Somer's D is used to determine the strength and direction of relation between pairs of variables. Its values range from -1.0 (all pairs disagree) to 1.0 (all pairs agree). It is defined as $(n_c-n_d)/t$ where n_c is the number of pairs that are concordant, and n_d the number of pairs that are discordant, and t is the number of total number of pairs with different responses.

f. **Gamma** - The Goodman-Kruskal Gamma method does not penalize for ties on either variable. Its values range from -1.0 (no association) to 1.0 (perfect association). Because it does not penalize for ties, its value will generally be greater than the values for Somer's D.

g. **Tau-a -** Kendall's Tau-a is a modification of Somer's D to take into the account the difference between the number of possible paired observations and the number of paired observations with different response. It is defined to be the ratio of the difference between the number of concordant pairs and the number of discordant pairs to the number of possible pairs $(2(n_c-n_d)/(N(N-1))$. Usually Tau-a is much smaller than Somer's D since there would be many paired observations with the same response.

h. **c -** Another measure of rank correlation of ordinal variables. It ranges from 0 to (no association) to 1 (perfect association). It is a variant of Somer's D index.

[4] Source: http://www.ats.ucla.edu/STAT/sas/output/sas_ologit_output.htm

The SURVEYLOGISTIC Procedure

Model Information	
Data Set	WORK.CENSUSSIMDQ_ED
Response Variable	response
Number of Response Levels	2
Stratum Variable	simdqui
Number of Strata	5
Cluster Variable	cluster_SIMD
Number of Clusters	20
Model	Binary Logit
Optimization Technique	Fisher's Scoring
Variance Adjustment	Degrees of Freedom (DF)

Response Profile		
Ordered Value	**response**	**Total Frequency**
1	0	28,326
2	1	22,775

Probability modeled is response=1.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	70,238.843	67,625.980
SC	70,247.684	67,714.395
-2 Log L	70,236.843	67,605.980

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	2,630.8628	9	<.0001
Score	2,594.1136	9	<.0001
Wald	3,788.3530	9	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
cen_dist	4	254.2933	<.0001
simdqui	4	18.2930	<.0001
postout	1	241.9161	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.00169	0.0146	0.0134	0.9079
cen_dist	1	1	-0.4750	0.0468	102.8704	<.0001
cen_dist	2	1	-0.3967	0.0283	196.7861	<.0001
cen_dist	3	1	0.2321	0.0319	52.9341	<.0001
cen_dist	4	1	0.4507	0.0494	83.2700	<.0001
simdqui	1	1	-0.3795	0.0392	93.8540	<.0001
simdqui	2	1	-0.0808	0.0229	12.4978	0.0004
simdqui	3	1	-0.0319	0.0249	1.6461	0.1995
simdqui	4	1	0.1462	0.0202	52.4088	<.0001
postout	0	1	0.0482	0.0113	18.2930	<.0001

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
cen_dist 1 vs 5	0.515	0.447	0.593
cen_dist 2 vs 5	0.557	0.505	0.614
cen_dist 3 vs 5	1.044	0.942	1.157
cen_dist 4 vs 5	1.299	1.135	1.487
simdqui 1 vs 5	0.484	0.429	0.546
simdqui 2 vs 5	0.653	0.586	0.727
simdqui 3 vs 5	0.685	0.614	0.765
simdqui 4 vs 5	0.819	0.738	0.908
postout 0 vs 1	1.101	1.054	1.151

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	60.6	**Somers' D**	0.258
Percent Discordant	34.9	**Gamma**	0.270
Percent Tied	4.5	**Tau-a**	0.127
Pairs	645124650	**c**	0.629

Appendix D - SAS programs

D1 - Program 1: Modelling the Raw Census Test Counts, and Using Proc Logistic

```
/* Here r = number of forms received and n = number of forms delivered */

data Testeval.CENSUS_TEST_RAW;
input cen_dist postout income_asked r n;
datalines;
1     0     0     925    2880
1     0     1     959 2886
1     1     0     853   3011
1     1     1     892   3080
2     0     0     1579 3797
2     0     1     1503 3855
2     1     0     1442 3828
2     1     1     1493 3946
3     0     0     1489 2744
3     0     1     1394 2713
3     1     0     1537 2856
3     1     1     1408 2824
4     0     0     846   1461
4     0     1     1061 1669
4     1     0     924   1562
4     1     1     868   1519
5     0     0     894   1581
5     0     1     919 1638
5     1     0     848   1617
5     1     1     941   1634
;
run;


/* Model with all the design variables */

proc logistic data = Testeval.CENSUS_TEST_RAW descending;
class cen_dist postout income_asked / param=ref ref=last;
model r/n = cen_dist postout income_asked / scale = pearson rsq;
run;

/* Note that the “param=ref ref=last” argument specifies the baseline */
/* i.e. Breadalbane, postout, with income */


** Evidence to suggest that income is not significant ;
** So remove income ;

proc logistic data = Testeval.CENSUS_TEST_RAW descending;
class cen_dist postout income_asked / param=ref ref=last;
model r/n = cen_dist postout / scale = pearson rsq lackfit influence
iplots;
run;
```

```
*** Check if there is an interaction effect between census district and
method ;

proc logistic data = Testeval.CENSUS_TEST_RAW descending;
class cen_dist postout income_asked / param=ref ref=last;
model r/n = cen_dist postout cen_dist*postout/ scale = pearson rsq lackfit;
run;


** Single covariate models ;

proc logistic data = Testeval.CENSUS_TEST_RAW descending;
class cen_dist postout income_asked / param=ref ref=last;
model r/n = postout / scale = pearson rsq;
run;

proc logistic data = Testeval.CENSUS_TEST_RAW descending;
class cen_dist postout income_asked / param=ref ref=last;
model r/n = cen_dist / scale = pearson rsq;
run;

proc logistic data = Testeval.CENSUS_TEST_RAW descending;
class cen_dist postout income_asked / param=ref ref=last;
model r/n = income_asked / scale = pearson rsq;
run;
```

D2 - Program 2: Model Stratifying by the SIMD Deprivation Data and Using Proc Surveylogistic

```
/** First create SIMD deprivation deciles **/
/** There is a data set known as CENSUS_TEST_SIMD_ANL which contains    **/
/** the Census test data                                                **/

data work.censusSIMDD;
set testeval.CENSUS_TEST_SIMD_ANLv2;
if simdrank < 650 then simddec=1;
if 651 < simdrank < 1301 then  simddec=2;
if 1302 < simdrank < 1951 then  simddec=3;
if 1952 < simdrank < 2602 then  simddec=4;
if 2603 < simdrank < 3252 then  simddec=5;
if 3253< simdrank < 3903 then  simddec=6;
if 3904 < simdrank < 4553 then  simddec=7;
if 4554 < simdrank < 5204 then  simddec=8;
if 5205 < simdrank < 5854 then  simddec=9;
else if simdrank > 5855 then  simddec=10;
run;


proc freq data=work.CensusSIMDD;
      tables cen_dist*income_asked*postout*simddec*response / nocol norow
nopercent ;
 run;


/** Because of data paucity it is better to create SIMD deprivation
quintiles **/

data work.censusSIMDQ;
set testeval.CENSUS_TEST_SIMD_ANLv2;
if simdrank < 1301 then simdqui=1;
if 1302 < simdrank < 2602 then  simdqui=2;
if 2603 < simdrank < 3903 then  simdqui=3;
if 3904 < simdrank < 5204 then  simdqui=4;
else if simdrank > 5204 then simdqui=5;
run;

proc freq data=work.CensusSIMDQ;
      tables cen_dist*income_asked*postout*simdqui*response / nocol norow
nopercent ;
 run;
 /** Use "surveylogistic SAS procedure" to perform analysis **/
 ** This is allows us to include the survey design methodology in analysis;
 ** First use the 'strata' subcommand,  for the quintiles ;


** Initially model the saturated model, with all variables;

proc surveylogistic data=work.censusSIMDQ;
 strata simdqui;
 class cen_dist income_asked postout simdqui;
model response (event='1')= cen_dist income_asked postout
cen_dist*income_asked cen_dist*postout
            income_asked*postout cen_dist*income_asked*postout;
run;


** There's evidence to suggest that 'INCOME' is not significant;
** But to fully understand need to model the factors individually - i.e.
investigate the main effects;

proc surveylogistic data=work.censusSIMDQ;
 strata simdqui;
 class cen_dist income_asked postout simdqui;
model response (event='1')= cen_dist income_asked postout;
run;


** Income not significant, so remove income;

proc surveylogistic data=work.censusSIMDQ;
 strata simdqui;
 class cen_dist income_asked postout simdqui;
model response (event='1')= cen_dist postout;
run;


** Investigation of interaction effects;

proc surveylogistic data=work.censusSIMDQ;
 strata simdqui;
 class cen_dist income_asked postout simdqui;
model response (event='1')= cen_dist postout cen_dist*postout;
run;


** evidence to suggest that the interaction between the census district and
postout is not significant;


*** CHECK WHAT HAPPENS WHEN SIMDQUI IS INCLUDED AS A PARAMETER IN THE
MODEL;

proc surveylogistic data=work.censusSIMDQ;
 strata simdqui;
 class cen_dist income_asked postout simdqui;
model response (event='1')= cen_dist postout simdqui;
run;
```

```
proc surveylogistic data=work.censusSIMDQ;
 strata simdqui;
 class cen_dist income_asked postout simdqui;
model response (event='1')= cen_dist postout simdqui;
run;


/** LASTLY CHECK THE SINGLE DESIGN VARIABLE MODELS */

proc surveylogistic data=work.censusSIMDQ;
strata simdqui;
class cen_dist income_asked postout;
model response (event='1')= cen_dist;
run;

proc surveylogistic data=work.censusSIMDQ;
strata simdqui;
class cen_dist income_asked postout;
model response (event='1')= postout;
run;

proc surveylogistic data=work.censusSIMDQ;
strata simdqui;
class cen_dist income_asked postout;
model response (event='1')= income_asked;
run;
```

D3 - Program 3: Final Model which Stratifies by the SIMD and Investigates Whether there is Clustering

```
* Investigation of the enumeration district effects;

proc surveylogistic data=testeval.CENSUS_TEST_SIMD_ALL_ANLv2;
cluster cluster;
strata unique_ed;
class cen_dist income_asked postout cluster_ED;
model response (event='1')=cen_dist income_asked postout;
run;


*** Results show that 'income' is not significant;
** So need to remove 'income';

proc surveylogistic data=testeval.CENSUS_TEST_SIMD_ALL_ANLv2;
cluster cluster_ED;
strata unique_ed;
class cen_dist income_asked postout cluster_ED;
model response (event='1')=cen_dist postout;
run;


*** Stratify by SIMD;
*** Create SIMD quintiles;

data work.censusSIMDQ_ED;
set testeval.CENSUS_TEST_SIMD_ALL_ANLv2;
if simdrank < 1301 then simdqui=1;
if 1302 < simdrank < 2602 then  simdqui=2;
if 2603 < simdrank < 3903 then  simdqui=3;
if 3904 < simdrank < 5204 then  simdqui=4;
else if simdrank > 5204 then simdqui=5;
run;


/* Model with both ED strata and clusters */

proc surveylogistic data=work.censusSIMDQ_ED;
cluster cluster_ED;
strata ed;
class cen_dist income_asked postout ed simdqui;
model response (event='1')=cen_dist postout ed;
run;


/*** As expected adding the ‘ED’ as a covariate in the model improves the
**** model fit; in that the model’s predictive power increases.
**** However, the simpler model without ED is preferred due to parsimony.*/
```

```
*** Creation of new cluster variable, based on SIMD quintiles ;
*** This is done by concatenation of the unique_ED, income, postout and
*** simdqui variables ;
*** A new data set in the WORK library is created ;

data work.censusSIMDQ_ED2;
set work.censusSIMDQ_ED;
cluster_SIMD=cat(of unique_ed income_asked postout simdqui);
run;


** Run logistic model using the 'cluster' variable and SIMD quintiles as
strata;

proc surveylogistic data=work.censusSIMDQ_ED2;
cluster cluster_SIMD;
strata simdqui;
class cen_dist income_asked postout simdqui;
model response (event='1') = cen_dist income_asked postout;
run;


** 'income_asked' not significant ;
* REMOVE  income ;

proc surveylogistic data=work.censusSIMDQ_ED2;
cluster cluster_SIMD;
strata simdqui;
class cen_dist income_asked postout simdqui;
model response (event='1') = cen_dist postout;
run;


*** Check the univariate models ;

proc surveylogistic data=work.censusSIMDQ_ED2;
cluster cluster_SIMD;
strata simdqui;
class cen_dist income_asked postout simdqui;
model response (event='1') = postout;
run;

proc surveylogistic data=work.censusSIMDQ_ED2;
cluster cluster_SIMD;
strata simdqui;
class cen_dist income_asked postout simdqui;
model response (event='1') = income_asked;
run;
```

```
****** POSTOUT IS NOT SIGNIFICANT (MARGINALLY), I.E. P-VALUE IS 0.0642 ;

proc surveylogistic data=work.censusSIMDQ_ED2;
cluster cluster_SIMD;
strata simdqui;
class cen_dist income_asked postout simdqui;
model response (event='1') = cen_dist;
run;


/**************************************************************************
* Conclusion: Whichever way the analysis proceeds the income question     *
* is shown not to have a significant on the propensity to respond.        *
* The preferred model stratifies by SIMD, with suitable cluster effects   *
*                                                                         *
**************************************************************************/
```

Appendix E - Glossary of terms

Binary Response

A response which is one of two states. For example, “yes” or “no”.

Blocking

The arranging of experimental units in groups (blocks) that are similar to one another.

Categorical Data

A set of data which can be sorted into distinct categories. For example, people have the characteristic of “gender” with categories “male” and “female”.

Clustering

Clustering exists where “natural” groupings are evident For example, neighbouring households can be regarded as a cluster.

Confidence Interval

A confidence interval (CI) is an interval estimate of a population parameter. Instead of estimating the parameter by a single value, an interval likely to include the parameter is given. Thus, confidence intervals are used to indicate the reliability of an estimate.

Correlation

Correlation indicates the strength and direction of a relationship between two variables. For example, households/individuals with similar characteristics may respond similarly to certain questions. See also regression.

Exp

The exponential function in mathematics.

Explanatory Variable

Explanatory variables are those which help explain a change in a response variable. For example, enumeration method (explanatory) may help explain response rates (response). Also known as independent variable. See also response variable.

Exploratory Data

Exploratory data is required in any applied statistical analysis to get a feel for the data to allow for the formulation of hypotheses worth testing.

Factor

A variable which is deliberately varied between trials in order to study its influence on the outcome.

Hypothesis

A prediction or solution to a problem that is typically written as a question. For example, enumeration strategy has an effect on census response rates.

Interaction

Interaction occurs where two or more objects, or variables, have an effect upon one another. For example, interaction can occur between a driver and the position of their car on a road.

Intra-cluster Correlation

Units in the same cluster are likely to be more similar than two units picked at random. For example, neighbouring households in one town maybe more similar than two households from separate areas.

Log
The logarithm function in mathematics.

Logistic Regression
A model used for prediction of the probability of occurrence of an event. It makes use of several predictor variables. For example, we can predict the probability of a household responding to the 2006 Census Test using predictor variables such as enumeration method, and/or the inclusion of income question.

Odds Ratio
The odds ratio is defined as the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. For example we could compare the odds (probability) of a household responding to the census when their questionnaire has the income question included, against a household questionnaire which does not have the income question included. An odds ratio of 1 indicates that the event is equally likely in both groups.

P-value
The probability (ranging from zero to one) that the results observed in a sample could have occurred by chance. The smaller the p-value, the more evidence we have the result is fair. Statistical analysis typically looks for p-values to be less than 0.05 to indicate a result is significant.

Probability
The likelihood or chance that something will happen, or has happened. For example, the probability that an individual is male or female.

Proxy Measures
A measure from which a variable of interest can be obtained or derived. For example, country of origin or birthplace might be used as a proxy for ethnicity.

Purposive Sampling
Choosing a sample based on who would be appropriate for the study, rather than selecting at random.

Randomisation
The process of making something random. For example, randomly selecting an area within Scotland as a suitable sample for testing certain questions.

Regression
A technique used for determining and/or modelling a relationship between two or more variables. For example, we might consider the height and weight of a sample of adults. See also correlation.

Response Variable
Those variables that change in response to an explanatory variable. For example, the weight of a person (response) may be affected by the height of a person (explanatory). Also know as dependent variable. See also explanatory variable.

Sampling (Sample)
The use of a subset of a population to yield some knowledge about a population of concern. For example, collecting information within chosen geographical areas as being a representative sample of your population as a whole.

SAS SAS (pronounced “sass”, originally Statistical Analytical System) is an integrated system of software products allowing a vast range of data and statistical analysis to be performed.

Standard Error Standard errors provide simple measures of uncertainty in a value. For example, in surveys reporting public attitudes, the larger the standard error, the less confidence one should have that the reported figures are close to the true values.

Stratification The process of grouping members of a population into relatively equal subgroups before sampling. For example, stratifying all households within Scotland into subgroups defined as the Local Authority area.

Variance (Variation) A measure of statistical dispersion, which captures the scale or degree of data being spread out. For example, a large variance in a response variable such as age indicates a large spread of responses.