ONS(ONC(SC))97/10

ONE NUMBER CENSUS STEERING COMMITTEE

Design of the Census Coverage Survey

1. The attached paper reviews the research into the proposed design for the Census Coverage Survey following the 2001 Census part of a One Number Census.

2. Further work is planned to:

   a) investigate the design for groups of counties;

   b) calculate sample sizes for all counties;

   c) investigate the impact of a large number of zero counts on the regression estimator;

   d) compare regression and dual-system estimators.

3. The Steering Committee are asked to:

   a) note the paper

   b) provide any comments (at the forthcoming meeting or in writing by 10 December 1997) on the proposed plans for further research.

Lisa Buckner
Census Division
Office for National Statistics

Room 4200W
Segensworth Road
Titchfield
Fareham
Hampshire PO15 5RR

November 1997

Design of the Census Coverage Survey

James Brown, Ian Diamond, Ray Chambers, Lisa Buckner

1. Introduction

1.1 This paper outlines a proposed strategy for a Census Coverage Survey (CCS) to follow the 2001 Census. A model based approach is adopted for the design and direct estimation. This allows full advantage to be taken of the highly correlated auxiliary information available after the 2001 Census. The aim is to estimate underenumeration at a County (or group of counties) level (by age and sex); allocation of this underenumeration down to lower levels is considered in the small area adjustments paper (ONS(ONC(SC))97/13).

A simulation study is being undertaken to assess the design and direct estimation procedures. A description and initial results from this are included in this paper.

Coverage versus validation

1.2 Following the 1991 UK Census a Census Validation Survey (CVS) was carried out in England, Scotland, and Wales. The survey aimed:

a) to estimate net underenumeration; and

b) to validate the quality of Census data.

1.3 The second aim required the re-enumeration of a sample using the entire Census form. This requirement is costly, due to time, resulting in a small sample size. The small sample size contributed to the fact that the survey failed in its first aim on two points: the survey failed to find, at a national level, anywhere near the number of people that the demographic rolled forward estimate suggested were missing and it was unable to estimate the geographic variation of underenumeration. It is proposed that the survey in 2001 should address coverage exclusively. Information on the quality of Census data would be obtained from Census Tests in 1997 and 1999. This allows for a much shorter doorstep questionnaire. Savings in time can be translated into more sample units.

2. A postcode based survey

2.1 The proposal is for a postcode-unit based survey. This requires the re-enumeration of a sample of postcode units rather than households. This clustering of the sample permits a larger sample size. While that does not necessarily improve the direct estimation due to clustering effects, it is important for estimating adjustments at lower levels.

2.2 The sample of postcodes needs to be sufficiently large to estimate the total population by 24 age-sex groups, for each design level group[1]. At the design level, postcodes are stratified into groups by a 'Hard to Count' (HtC) index and size. In this paper, 'hard to count' is defined in terms of characteristics found to be important after the 1991 Census by ONS and the Estimating With Confidence Project (Simpson *et al.*, 1997). The problem is to estimate 24 age-sex totals such that each has a Relative Standard Error (RSE) of less than a certain percent at the design level.

2.3 In general, postcode level information, beyond number of addresses, is not known. This leads to a two-stage design, selecting enumeration districts as Primary Sampling Units (PSUs) and then

[1] Each design level group is either a single county or group of smaller contiguous counties.

sampling postcodes as Secondary Sampling Units (SSUs) within selected enumeration districts. Clustering from the two-stage design has cost advantages for a fixed number of postcodes but efficiency disadvantages when the characteristics of postcodes are positively correlated within enumeration districts.

3. Direct estimation from the Census Coverage Survey

3.1 The quantities of interest are:

Z_{aidc} = 1991 adjusted Census count for age-sex group a of postcode i, in hard to count group d of county c.
X_{aidc} = 2001 unadjusted Census count.
Y_{aidc} = True 2001 count (given by the CCS for those postcodes in sample).

where:

c = 1...C design level county groups in England & Wales.
d = 1...D hard to count categories of postcodes.
a = 1...24 age-sex groups (0-4, 5-9, ..., 40-44, 45-79, 80-84, 85+).
i = 1...N_{dc} postcodes in hard to count group d of county c of which n_{dc} are in the sample S, the rest are in the non-sample R.

3.2 For direct estimation from the CCS it is required that the total populations T_{ac} be estimated to a certain degree of accuracy. This is treated as 24 similar estimations within each design level group. For this reason the design and estimation for one age-sex by design level group is described below. The same model framework applies for all other age-sex groups and in the following the subscripts a and c are dropped.

4. Stage One of the CCS design

4.1 A robust non-parametric model for stage one is a stratified super-population model of enumeration districts with simple random sampling within each stratum. Within a design level group the enumeration districts are stratified by hard to count. This is important as within the design group, undercount will depend on the characteristics of the PSUs. It also ensures that the CCS sample is spread across the full range of enumeration districts. Further stratification by size using the 1991 adjusted Census counts improves efficiency by reducing within stratum variance. Ideally one would like to use the 2001 counts but the CCS must be ready for the field directly after the Census so this is not possible. It is expected that the final design will use 1991 based estimates of the population in 2001.

4.2 Allowing for h = 1...H_d size strata within each hard to count group, *and in this case using i for enumeration districts rather than postcodes*, the model for a given age-sex group within a design level group can be written as:

$$\left.\begin{aligned} E_\xi\{Y_{ihd}\} &= \mu_{hd} \\ Var_\xi\{Y_{ihd}\} &= \sigma^2_{hd} \end{aligned}\right\} \; i \in \text{ h within d}$$

$$Cov_\xi\{Y_i\ ,\ Y_j | X_i\ ,\ X_i\} = 0 \text{ for all } i \neq j$$

4.3 Assuming no second stage sample, estimation of the required total is straightforward under this model using a stratum by stratum expansion estimator. From this it is possible to calculate the sample of enumeration districts required if there was no second stage sample.

5. Stage Two of the CCS design

5.1 It is possible to write-down a model for the second stage. This is more complicated due to the varying numbers of postcodes within enumeration districts, expected but unknown correlation of postcodes within enumeration districts, and the absence of readily available postcode level information on which to design. Therefore, the proposal is to have a constant second stage sample size and take a simple random sample of postcodes within chosen enumeration districts. This has the appeal of simplicity in the absence of detailed postcode information. Any loss of efficiency due to the second stage sample is examined more fully through the simulation study presented in Section 9.

The CCS super-population model for estimation

5.2 It is sensible to assume that the 2001 Census count and the CCS count within each postcode will be related. If this is not true then one really should be suspicious of one of the counts. Further, within sub-groups of postcodes a linear relationship may well be appropriate. This corresponds to a constant ratio (or adjustment factor) between the two counts with the possibility of a non-zero constant. The constant is needed in some postcodes where the Census misses *all* people from a certain age-sex group. It is the possible occurrence of this situation which leads to choosing the Regression Estimator in preference to the Ratio Estimator which forces the constant to be zero. Given that it is know from the 1991 Census that age and sex are crucial to undercount, as well as local characteristics, it is sensible to consider a model within age-sex groups for each hard to count by design level group where the hard to count index allows for different local characteristics. The simple regression model stratified by the hard to count index for an age-sex group is:

$$\left.\begin{aligned} E_\xi\left\{Y_{id} \middle| X_{id}\right\} &= \alpha_d + \beta_d X_{id} \\ Var_\xi\left\{Y_{id} \middle| X_{id}\right\} &= \sigma_d^2 \end{aligned}\right\} \; i \in d$$

$$Cov_\xi\left\{Y_i\,,\, Y_j \middle| X_i\,,\, X_j\right\} = 0 \qquad \text{for all } i \neq j$$

5.3 Substituting in the Ordinary Least Squares (OLS) estimators for α_d and β_d it is straightforward to show (Royall, 1970) that under this model the Best Linear Unbiased Predictor (BLUP) for the total T is:

$$\hat{T}_\xi = \sum\nolimits_d \left\{ T_{Sd} + \sum\nolimits_{Rd} \left(\hat{\alpha}_d + \hat{\beta}_d X_{id} \right) \right\} \quad (1)$$

where $\sum_{Rd}$ is the summation over all non-sample postcodes in hard to count index d and $\sum_d$ is the summation over all the hard to count groups. Strictly speaking the model is known to be wrong (the postcodes are correlated within enumeration districts), but the simple two stage model proposed by Scott and Holt (1982), which assumes independence between PSUs, is still reasonable. Under this model they state that the OLS approach remains unbiased and the loss of efficiency is negligible as the residual correlation within clusters, that is the correlation left unexplained, tends to zero. (In practice the residual correlation is much less than the correlation in the data before fitting the model and it is very close to zero.)

5.4 There is a model-based formula for estimating the variance of $\hat{T}_{\xi} - T$, the estimation error, under the model. Unlike the estimator of the total, this is sensitive to mis-specification of the variance structure even when the design is *approximately* balanced with respect to the auxiliary variable (Royall and Cumberland, 1978). In this strategy it is proposed that the conservative ultimate cluster variance estimator, a variant of the random groups approach, be used as the postcodes are clustered within enumeration districts. Once the variances are estimated an estimated RSE can be calculated for each age-sex group total. In general, when the regression model is appropriate, the estimator in (1) is more efficient than a simple stratum by stratum expansion estimator for a given sample size.

6. Case Study: Applying Stage One of the CCS Design to Hampshire

6.1 Hampshire was chosen purely for convenience to examine the feasibility of Stage One of the design. It was considered to be an 'average' county with just over 3,000 enumeration districts and includes two middle-sized cities. Some counties are considerably smaller hence the need in some cases to group contiguous counties at the design level.

National hard to count index

6.2 The first stage was to calculate a national hard to count index for SASPAC[2] enumeration districts based on the 1991 Census. Only enumeration districts with non zero populations in 1991 were used. The variables used were percent of permanent households which are private rented, percent of residents imputed and percent of young residents who were in-migrants in the last year (from SASPAC) with percent of households in multi-occupied buildings and percent of residents with language difficulty provided by the Office for National Statistics (ONS). Increasing values for any of these variables are expected to increase undercount. To form an index the normalised ranks of each variable were added and the sum split into quintiles, 1 being the easiest to count and 5 being the hardest to count enumeration districts. Using the normalised ranks avoids outliers, for any one component variable having undue weight in an enumeration district's score on the index.

6.3 Within Hampshire there are 3,305 enumeration districts on SASPAC of which 3,229 had a non zero population in the 1991 Census and therefore an index value. The distribution of the districts by the index is given in Table 1 below.

[2] SASPAC is the package used by Census users to access small area statistics. A few enumeration districts are excluded from SASPAC by ONS due to there small size although enumeration districts with zero counts are included.

Table 1. Distribution of 1991 Hampshire enumeration districts by hard to count index

Hardness To Count	Number of Enumeration Districts (Current HtC Index)	Number of Enumeration Districts (Previous HtC Index)
Very Easy	249	892
Easy	626	717
Medium	874	678
Hard	925	601
Very Hard	555	341

6.4 The distribution in column 2 of Table 1 reflects the presence of Portsmouth and Southampton, two cities on the south coast, in Hampshire with a predominance of the harder to count groups. Table 1 also presents the distribution of enumeration districts based on the HtC index used in paper ONS(ONC(SC))97/02. This earlier index concentrated on the social characteristics of areas whereas the current index attempts to take some account of the practical barriers, such as multiple occupancy, to obtaining a Census response.

Multivariate stratification and design

6.5 Within a hard to count by design level group, estimation is required for each age-sex group, therefore there are 24 potential size variables, each of the Z_{ai}'s, to stratify on. The solution adopted here is a multivariate approach that uses six key age-sex groups, male and female 0-4, males 20-34, and females 85+. The choice of these is based on a coverage analysis of the 1991 Census. 28 large enumeration districts, based on counts of males aged 20-34, were excluded from the rest of the design process and treated as outliers. These were included in the final design with probability 1. Principal Component Analysis, was used on the remaining 3201 districts. The first three component scores, which accounted for over 96 percent of the original variability, were kept. Within index group Ward Linkage Cluster Analysis on these components was then used to form minimum variance strata. At this stage, additional enumeration districts were highlighted as outliers and included in the sample with probability 1 and excluded from the sample size calculations.

6.6 To calculate the total sample size a design variable Z_i based on the chosen principal components was constructed from the formula:

$$Z_i' = \frac{|V| \times \sum_j P_{ji}}{\left\{\sum_j \text{var}(P_{ji})\right\}^{1/2}} \qquad (2)$$

where $\sum_j$ is over the principal components chosen, P_{ji} is the j^{th} component score for the i^{th} enumeration district, and V is the variance-covariance matrix of the six original size variables. Using the determinant of this matrix as a measure of variability in the original data, and bearing in mind that principal components are orthogonal, the variance of the design variable in (2) is scaled to something which represents the original variability in all the variables. It should be noted that this multivariate approach is a change from the one presented in paper ONS(ONC(SC))97/02. The original intention was to include all 24 age-sex totals in the stratification. However, further work revealed that the original design variable based on all 24 age-sex groups and only the first two principal components did not sufficiently reflect the true variability for any age-sex groups. When this was corrected for, the sample size obtained was $^1/_3$ of the total population. The current

procedure is a compromise that concentrates the efficiency gains from stratification in the key age-sex groups.

6.7 This design variable was then used to calculate the total sample required for an RSE of one percent[3] with respect to the design variable. (0.5 was not practical once the design was correctly accounting for the variability in the age-sex counts as it led to a sampling fraction at Stage One of 20 percent.) Annex A sets out the theory in detail giving the appropriate formula for the super-population model specified in Section 5.2. Optimal allocation, also known as Neyman allocation, was used to allocate the sample to the strata with the condition that the minimum sample was one enumeration district. This constraint raises the final sample size to which the outliers were added.

Final sample design

6.8 Several different size stratifications were tried by varying the number of clusters required in the clustering algorithm. In general, increasing the number brings down the total sample size however, it also increases the number of outliers identified by the algorithm and all clusters have a sample of at least one. This means there is a trade-off between the two and after a point increasing the number of size strata does not improve the design. The final design and allocation is given in Table 2 below.

Table 2 - Sample allocation for the first stage sample in Hampshire

Index Group	Population Size	Number of Size Strata	Sample Size	Outliers[b]
Very Hard	246	15	27	3
Hard	623	35	59	1
Medium	863	35	80	2
Easy	918	35	86	3
Very Easy	551	30	56	2
Outliers[a]	28	-	28	-
TOTAL	3229	150	336	11

a. Enumeration districts classified as outliers based on their size.
b. Enumeration districts classified as outliers by the clustering algorithm.

6.9 From Table 2 it would appear that more size strata would further reduce the sample but the gains are small and these are countered by more outliers of type b. This increasing of outliers as a result of requiring more clusters may be reduced by trying other clustering algorithms. However, this analysis has not yet been carried-out. Further work to identify the characteristics of outlying enumeration districts will also be necessary when the final design is calculated for all county groups.

6.10 The design in Table 2 gives a total first stage sample of 347 enumeration districts, approximately a 10 percent sampling fraction. This is with respect to the design variable given by (2). To assess how well the design works for each individual age-sex variable the expected RSEs were calculated based on the design for the 3190 enumeration districts not classified as outliers and taking a sample of 308. These ranged from 1.4 percent for the 0-4 males to 4.6 percent for the 85+

[3] An RSE of percent for total T translates into an approximate 95 percent confidence interval on T of ±2 percent.

males. The six age groups in the design variable all had expected RSEs of less than 1.7 percent. As the RSE put on the design variable was one percent this shows that the design variable is not perfect. There is also extra loss of efficiency from taking the second stage sample. This will be balanced by the extra efficiency of the regression estimator. The extra efficiency of the regression estimator comes from using the auxiliary information. This means that instead of looking at the marginal variance of the CCS count within a hard to count group, the variance conditional on knowing the Census count is only considered. The effect of this in practice is examined in later sections of the paper using the simulation study.

7. Conclusions about the design

7.1 The design proposed here for the first stage is a standard approach. The auxiliary information is used to stratify, a standard procedure in both the model-based and design-based frameworks for making efficiency gains. The estimation model is chosen to make further efficiency gains using the additional auxiliary information available from the 2001 Census. These gains are related to the variability in Census coverage as this affects the conditional variance in the model. However, the conditional variance will always be less than the marginal variance when a regression model is sensible, leading to some efficiency gain. The case study for Hampshire deals with the practical application of the design. It shows that the theoretical framework proposed can be applied to an actual county with feasible results. However, Table 2 does not represent the final design for the 2001 CCS in Hampshire. In the final design it is likely that the Isle of Wight will be included in a group with Hampshire.

8. Extension to National Sample Size

8.1 At this stage it is useful to get a feel for the kinds of sample sizes that may be required may be required nationally (England and Wales) for a range of design RSEs. For Stage One the sampling unit is the enumeration district. Counties are very variable in the number of units they contain while heterogeneity amongst the units within counties is more similar. Unless county size is controlled for this will lead to very high sampling fractions for smaller counties. This is the reason for suggesting grouping contiguous counties to make pseudo counties, similar in size to Hampshire, of about 3000 enumeration districts or approximately 1.5 million people. An initial grouping has been made which reduces the 55 England and Wales counties to 34 groups. This grouping also accounts for splitting Inner London, Outer London, Greater Manchester, and West Midlands as these are much larger than 3000 enumeration districts.

8.2 The design has been implemented in Kent and West Yorkshire for a range of RSEs. These counties were chosen as Kent is approximately 3,000 enumeration districts, the average size required, and West Yorkshire is the largest single county which has not been split. The results are in Table 3.

Table 3 - Sample allocation for the first stage sample in Kent and West Yorkshire

RSE	Strata	Sample	Outliers	Total Sample
		KENT - 3158 EDs		
1.0	190	268	43	311
1.5	122	162	36	198
2.0	105	123	31	154
		WEST YORKSHIRE - 4098 EDs		
1.0	125	314	36	350
1.5	100	171	33	204
2.0	100	122	33	155

8.3 These two counties cover 7,256 enumeration districts out of approximately 110,000. With caution one can extrapolate to national sample sizes and get approximate figures of:

- 40,000 postcodes (approximately 600,000 households) for an RSE of 1.0 percent.
- 25,000 postcodes (approximately 375,000 households) for an RSE of 1.5 percent.
- 19,000 postcodes (approximately 245,000 households) for an RSE of 2.0 percent.

This extrapolation very much depends on Kent and West Yorkshire being a good representation of all county groups. They should achieve this for the 50,000 enumeration districts allocated to single county groups. However, one should question how well they represent the 60,000 enumeration districts in multiple county groups. Initial work with the group Bedfordshire, Buckinghamshire, and Northamptonshire suggests a slightly higher enumeration district sampling fraction for multiple groups, up to 50,000 postcodes for a 1 percent RSE. This is not unexpected when county is being imposed as an extra layer of stratification. These preliminary results suggest more work is needed to find the 'best' way to implement the first stage of the design within groups of small counties.

9. Simulation Study of the Design and Estimation Procedures

9.1 The detail in Section 4 only really examines the first stage of the design. The aim of this simulation study is to examine the performance of the design when the Second Stage sample is taken. It is also necessary to see how appropriate and efficient the regression estimator is. This is particularly important as the expected RSEs in Annex C are larger than is required of such an important survey and to reduce these using the model framework of the design would require even larger samples. Extra efficiency is also needed since in 2001 the design will be based on 1991 data and will therefore be out-of-date.

Data used in the simulation

9.2 Anonymised individual records from the 1991 Census for one complete district from a county in England and Wales were used in the simulation. The Hard To Count Index was added to the data. The district is treated as a county in the design.

9.3 The county used has 445,351 individuals within 171,265 households. It consists of 11,330 postcodes (141 with only one person and 46 with over 200 people) and 930 enumeration districts (five have only one postcode, one has 40 postcodes, and the median is 14 postcodes). The distribution of enumeration districts by Hard To Count Index is given in Table 4 below.

Table 4 - Distribution of enumeration districts by hard to count index

Hardness To Count	Number of Enumeration Districts
Very Easy	144
Easy	210
Medium	186
Hard	193
Very Hard	197

9.4 The distribution in Table 4 is fairly even with respect to the index. This is a good test as it is necessary to avoid extremes, especially a situation where the very easy group dominates as this would tend to make the overall performance of the design too optimistic.

Simulation method used

9.5 Treating the data as the true, in reality unobservable, population the first stage of the simulation was to create a Census. Each individual was given a fixed probability of being counted in a Census based on their age, sex, and enumeration district hard to count index. This was done by simple random sampling with replacement from the population of Estimating With Confidence enumeration district adjustment factors. These are the 'best guess' at small area coverage for the 1991 Census. To create a Census, a binomial trial was carried-out for each individual. This was carried out independently and certain rules were then applied to ensure that counted households had a sensible structure. Households were excluded if:

- any children aged 5-15 were missed from a counted household
- all household members aged 16 and over were missed
- one partner from an elderly couple was missed.

This strategy for excluding households may not necessarily be a perfect representation of reality but presents a simple model. Also, characteristics of the undercount are fixed by the model at the start and therefore, as expected these are the characteristics found by the model as significant.

9.6 For the CCS, the design procedure used for Hampshire was followed but based on an RSE of 2.5 percent to reflect the smaller population of PSUs. The final design and allocation is given in Table 5 below.

Table 5 - Sample allocation for the first stage sample

Index Group	Population Size	Number of Size Strata	Sample Size	Outliers[b]
Very Hard	144	10	12	0
Hard	210	16	17	0
Medium	185	14	14	3
Easy	192	15	18	3
Very Easy	197	15	16	0
Outliers[a]	2	-	2	-
TOTAL	930	70	79	6

a. Enumeration districts classified as outliers based on their size.
b. Enumeration districts classified as outliers by the clustering algorithm.

9.7 The design in Table 5 was fixed throughout the simulation and used to get a total sample of 85 enumeration districts. A fixed sample of four postcodes (or the number of postcodes in the enumeration district if less than four) was taken at the Second Stage. For each sample the totals for each age sex group were estimated, the variances calculated using the ultimate cluster variance estimator and estimated RSEs calculated. Ideally, it would be desirable to simulate one CCS per Census as this most accurately reflects real life. Computationally Censuses are time consuming to simulate so a compromise of 10 CCSs for each of 100 Censuses was adopted.

Results

9.8 The mean total coverage across the 100 Censuses is 95 percent, the worst coverage by age sex group is males aged 20-24 with an average coverage of 89 percent. This is, in general, conservative for most counties compared to 1991. A few counties did do worse, particularly Inner and Outer London and those containing the large metropolitans districts (Heady *et al.,* 1994). The key measure of performance of the procedure is the estimated RSEs. The average estimated RSEs are given in Table 6 and based on 1000 CCSs.

9.10 Table 6 shows on average that the procedure does well and in all cases the average estimated RSE is better than the RSE predicted by the model framework used in the design. This shows that on average the regression estimator has enough extra efficiency over the design model to recover loss of efficiency from the second stage sample as well as bring down the RSE in those age groups not included in the multivariate part of the design. However, the standard errors do show that in most cases a significant percentage of CCSs still do worse than the design predicts and it cannot be guaranteed that the regression estimator will do better.

Table 6 Mean Relative Standard Errors for 1000 simulated CCSs on county

	Males				Females		
Age Group	Number of CCSs	Design RSE	Average[b] Estimated RSE	Age Group	Number of CCSs	Design RSE	Average[b] Estimated RSE
0-4	1000	2.73	2.16 (0.686)	0-4	1000	2.66	2.12 (0.614)
5-9	1000	3.86	2.41 (0.625)	5-9	1000	3.92	2.48 (1.212)
10-14	1000	4.52	2.34 (0.706)	10-14	1000	4.45	2.19 (0.696)
15-19	1000	4.45	2.26 (0.884)	15-19	1000	4.19	1.65 (0.544)
20-24	1000	3.33	2.53 (0.697)	20-24	1000	3.22	1.62 (0.598)
25-25	1000	3.02	2.38 (0.513)	25-25	1000	2.99	1.58 (0.383)
30-34	1000	2.92	2.04 (0.451)	30-34	1000	3.12	1.68 (0.364)
35-39	1000	3.94	1.87 (0.455)	35-39	1000	4.04	1.45 (0.377)
40-44	1000	4.18	1.42 (0.359)	40-44	1000	4.53	1.24 (0.392)
45-79	1000	2.83	0.48 (0.159)	45-79	1000	2.77	0.34 (0.129)
80-84	917[a]	7.67	2.17 (0.903)	80-84	999[a]	6.24	1.61 (0.530)
85+	659[a]	10.43	3.42 (1.749)	85+	1000	3.33	2.63 (0.850)

a. Calculation of the variance is not always possible due to zero postcode counts in the CCS.
b .The estimated standard deviation for the distribution of the RSE is given in brackets.

9.11 The estimated RSE is a good measure to use after the survey to assess performance provided that the estimator is unbiased and the variance estimator gives correct coverage. Both of these properties are looked at in Table 7. The bias is averaged over the 1000 CCSs. To assess variance coverage 95 percent confidence intervals are estimated using the estimated variance in each case. Table 7 reports the proportion of estimated intervals which contain the truth.

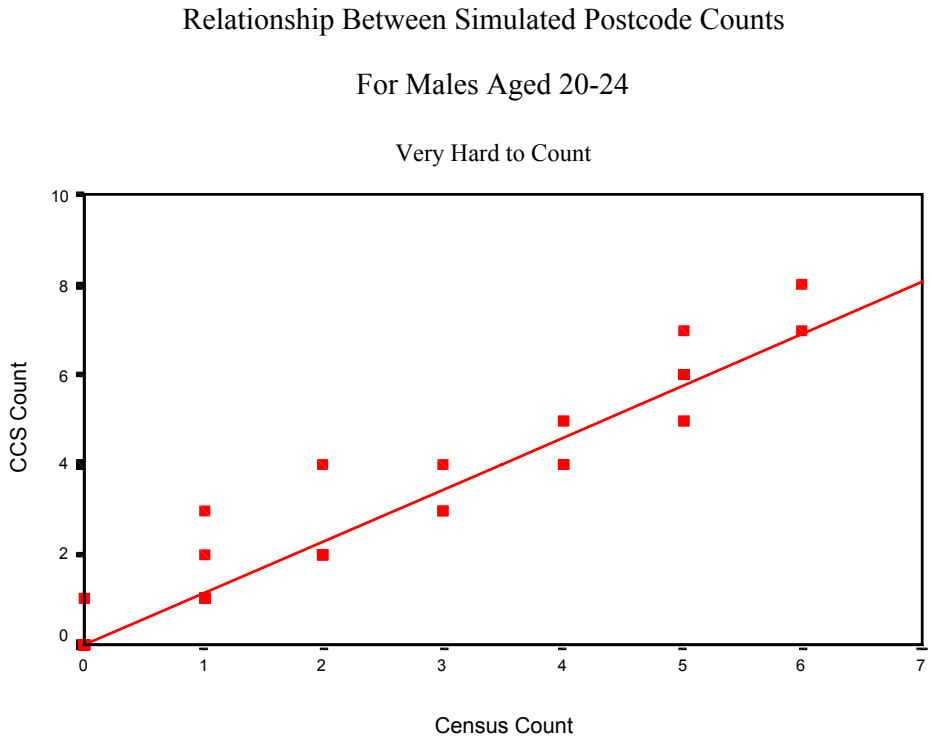
9.12 Table 7 shows that the variance estimator is giving good coverage with the 95 percent confidence interval containing the true value at least 95 percent of the time except in the two oldest age groups. Even in those cases the results are still good considering that variance estimation gets harder as the number of people in the age group decreases. One expects that the regression estimator would be unbiased over 1000 simulations if the regression model is appropriate so it is surprising that all age groups are showing a positive bias (over estimating the total). To check this further the relationship between the simulated Census and CCS counts was examined for one run of the simulation. Figure 1 is a scatter plot for a particular age sex group for enumeration districts from one level of the hard to count index.

Table 7 - Bias and Variance Coverage for 1000 simulated CCSs on county

	Males				Females		
Age Group	Number of CCSs	Bias[b]	Confidence Interval Coverage	Age Group	Number of CCSs	Bias[b]	Confidence Interval Coverage
0-4	1000	135.94	0.99	0-4	1000	108.27	0.99
5-9	1000	160.96	0.99	5-9	1000	118.35	0.99
10-14	1000	99.98	0.99	10-14	1000	129.64	0.98
15-19	1000	67.32	0.99	15-19	1000	58.76	0.97
20-24	1000	72.19	0.99	20-24	1000	53.13	0.99
25-25	1000	81.52	1.00	25-25	1000	69.16	0.99
30-34	1000	63.78	0.99	30-34	1000	131.49	0.99
35-39	1000	70.25	0.98	35-39	1000	48.45	0.97
40-44	1000	19.48	0.97	40-44	1000	8.98	0.95
45-79	1000	95.00	0.98	45-79	1000	37.83	0.97
80-84	917[a]	11.46	0.92	80-84	999[a]	1.83	0.91
85+	659[a]	1.24	0.95	85+	1000	18.45	0.90

a. Calculation of the variance is not always possible due to zero postcode counts in the CCS.
b. Calculation of the bias is averaged over all simulations including those for which the variance cannot be estimated.

Figure 1.



9.13 Figure 1 suggests that the regression model is appropriate. However, the fitted line is forced through the origin, not by the model as in the Ratio Model situation, but by the large number of postcodes that have zero counts for both the Census and CCS. These points not only pull the line

down but they also tend to rotate it. This rotation increases the gradient which is more influential in estimating the total than the constant and therefore causes the slight positive bias.

9.14 At this stage of the research the bias is noted as a potential problem which will require further investigation. However, overall the estimation model has done well and in relative terms the bias is very small as the total being estimated is of the order of 20,000.

10. Sensitivity analysis for Non-Perfect Dependent CCSs

10.1 The simulations so far have only considered a perfect independent CCS. Unfortunately, the real world is neither perfect or independent. The simulation program was extended to allow for non-response in the CCS and dependence between the Census and CCS. The dependence was achieved using a similar method to the one in ONS(ONC(SC))97/12 which involves varying the odds ratio between the Census and CCS. For a given odds ratio, a probability of being in the Census, and a probability of being in the CCS, the joint probabilities for all possible outcomes after the Census and CCS of an individual can be solved to complete the following 2x2 probability table:

	In CCS	Missed By CCS	
In Census	p_{11}	p_{10}	p_{1+}
Missed By Census	p_{01}	p_{00}	p_{0+}
	p_{+1}	p_{+0}	1

10.2 The values for overall Census coverage (p_{1+}) vary for each individual but do not vary across simulations. The values for the CCS response rate (p_{+1}) are fixed for each individual but vary across simulations from perfect (100%) to 95% and 80%. The odds ratio is varied from 0.1 (people not in Census are ten times more likely to be in the CCS than those counted in the Census) to 1 (independence) to 10 (people in Census are ten times more likely to be in the CCS than those not counted in the Census). This means that as the odds ratio decreases from one to zero the chance of finding different people in each increases (p_{11} and p_{00} go down). Conversely as the odds ratio increases from 1, p_{11} increases to its maximum value which is the smallest value of p_{1+} and p_{+1}.

10.3 The regression estimator was then used as before. However, once non-response was introduced into the CCS the Y count for a sample postcode in the model was taken to be the union of the Census and CCS count for that postcode. The assumption here is that p_{00} is zero. When this is violated it will introduce a negative bias into the estimate of the county totals.

Results of the sensitivity analysis

10.4 To compare the performance of the design for different levels of dependence the Relative Root Mean Square Error (RRMSE) was used. This is a fair measure of comparison as it accounts for variance and bias. The relative scale is used as the county totals being estimated are of the order of 20,000. The RRMSE is defined as:

$$\text{RRMSE} = \frac{1}{\text{truth}} \times \sqrt{\frac{1}{1000} \times \sum_{j=1}^{1000} \left(\text{observed}_i - \text{truth}\right)^2}$$

and is calculated within each age sex group across all 1000 simulations for each scenario. The relative bias is also considered and calculated as:

$$\text{Relative Bias } = \frac{1}{\text{truth}} \times \frac{1}{1000} \times \sum_{j=1}^{1000} \left(\text{observed}_i - \text{truth}\right)$$

for the same groups. The results are presented in a series of graphs for varying odds ratios by sex. Figures 2 to 4 are for males and show the RRMSE.

Figure 2.

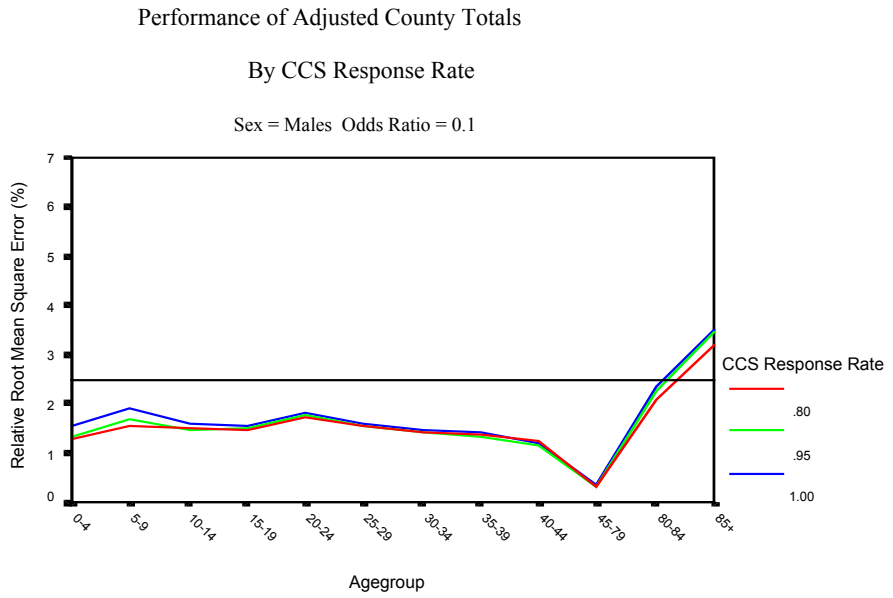


Figure 3.

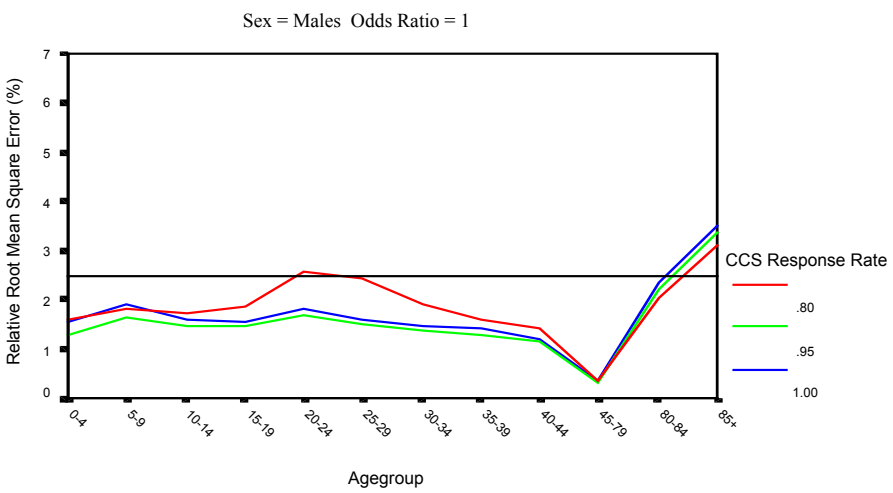
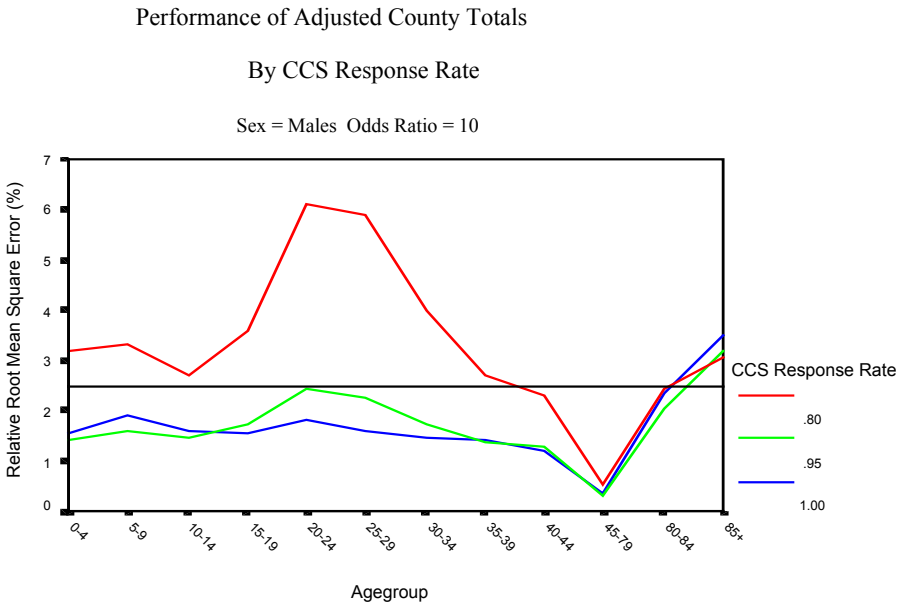


Figure 4.



10.5 Figures 2 and 3 show that for odds ratios of 0.1 and 1 even at an 80 percent response rate the RRMSE is still less than 2.5 percent. The exception is the oldest age group where the bias starts to dominate over the variance as the numbers being estimated decreases. However, as the odds ratio increases above one the same people tend to be missed by both to a greater extent. In this case, as the CCS response rate falls the RRMSE goes up, especially in those groups where the Census coverage is lower as well, such as males aged 20-29. The message here is that for a high CCS response rate the regression estimator will still do well regardless of dependence. As the CCS response rate falls, high levels of dependence in the wrong direction will lead to the regression estimator failing. In this case, 'wrong direction' means that is the odds ratio is greater than one and dependence between the Census and CCS results in missing the same people in both. The graphs for females are shown in Annex B and they display a similar pattern but to a lesser extent.

10.6 In this case the increasing RRMSE does not tell the whole story, it is necessary to see if this is being driven by the bias or the variance. Figures 5 to 7 present the relative biases for males for the changing odds ratios. For reference the relative bias for the unadjusted Census counts is also presented (No CCS line). This shows the effect of bias if no CCS is carried out after the Census.

10.7 Figures 5 to 7 show that as the RRMSE increases with the odds ratio the negative bias from the regression estimator is getting more important. Relative to the unadjusted counts the regression estimator still does very well for odds ratios of 0.1 and 1. However, for the odds ratio of 10, once the CCS response rate has fallen to 80 percent, the RRMSE is being entirely driven by the relative bias shown in Figure 7. It can be seen that by comparing the absolute relative bias in Figure 7 to the RRMSE in Figure 4 that the RRMSE is nearly all due to bias and not variance. This has serious consequences for calculation of confidence intervals from estimated variances as the confidence interval will be calculated around the wrong point. It should still be noted of course that even in this worst case you are still doing better than not adjusting at all. Again the graphs for females are shown in Annex B but the results again have the same but less extreme pattern.

Figure 5.

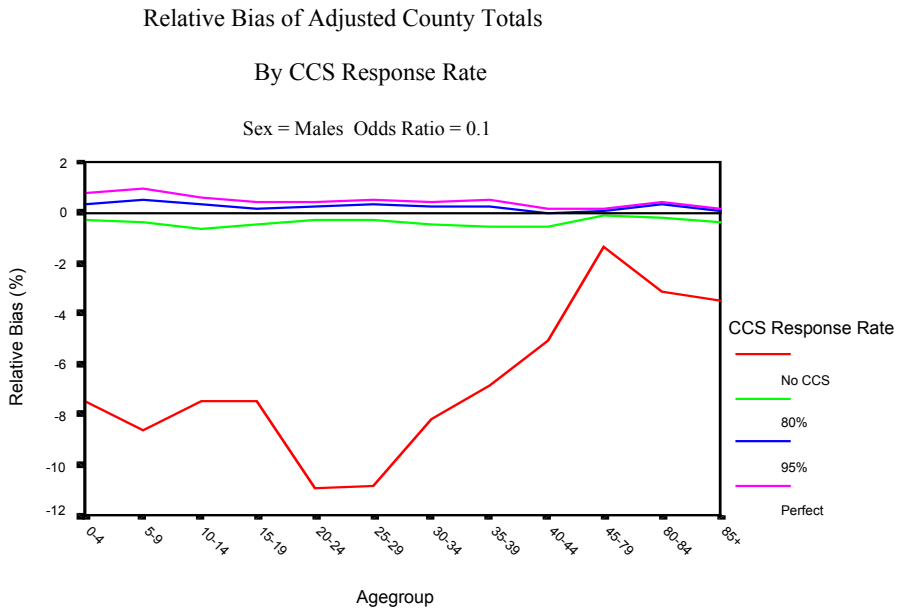


Figure 6.

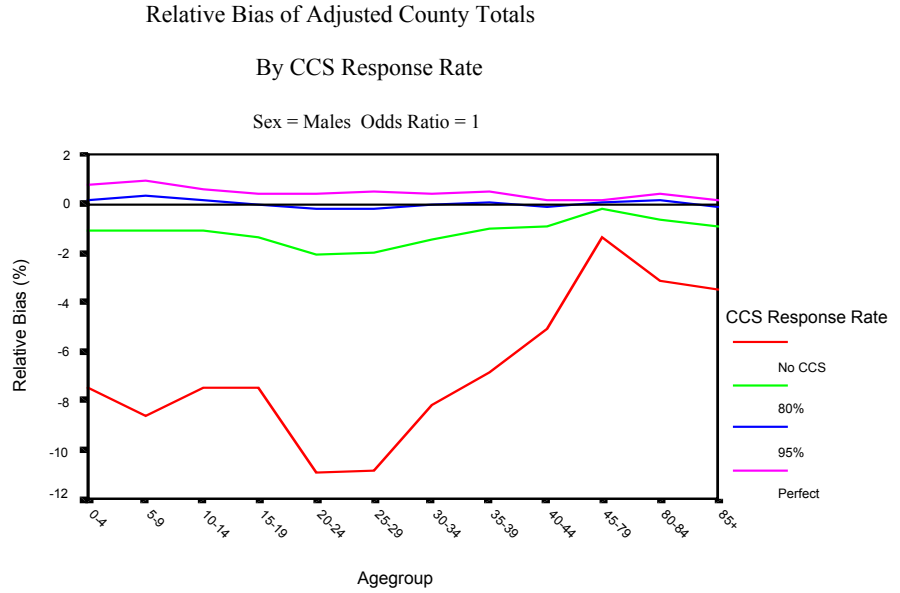
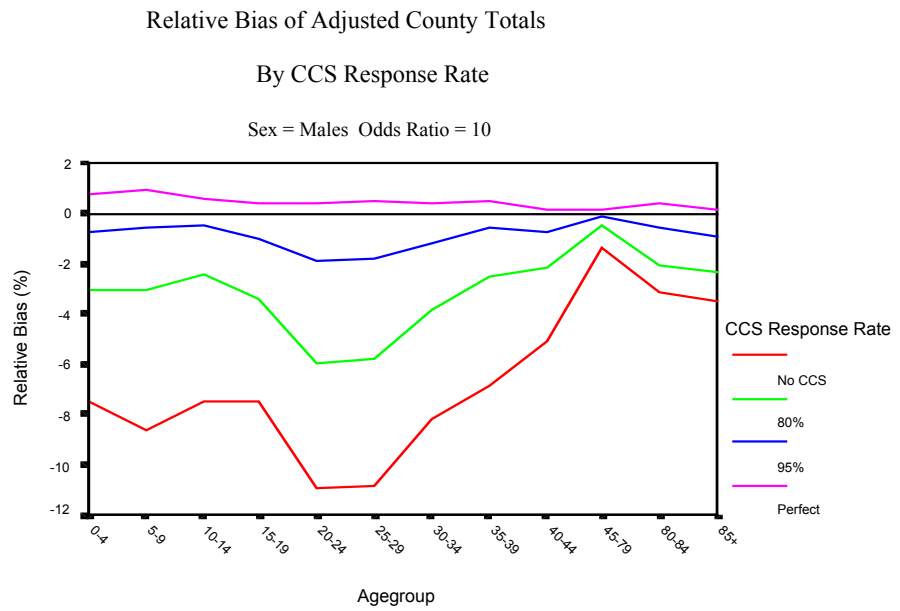


Figure 7.



10.8 From this initial sensitivity analysis it can be seen that determining the possible extent and direction of dependency between the Census and CCS is important. More work is needed to quantify how bad an odds ratio of 10 is, as it is only when the dependence is in this direction that there is cause for concern. As the odds ratio decreases to zero the regression estimator will not suffer, even if its response rate falls, as it will still find the different people for the union count.

11. Initial conclusions and further work

11.1 The initial conclusions are that for a perfect CCS the regression estimator is working well to recover any loss of efficiency due to the second stage design and multivariate stratification. However, the spread of RSEs is still quite high. The coverage of confidence intervals from the ultimate cluster variance estimator is excellent, even though there is a slight positive bias in the estimator due to zero count postcodes. As non-response is introduced with dependence the regression estimator still performs well. As the response rate for the CCS decreases the direction and the extent of the dependence becomes important, especially for those age-sex groups with the lowest Census response rates. This shows how important it is to get a high CCS response since once this is achieved the dependence becomes a side issue. High response is also important for variance estimation since for increasing odds ratios and low response the bias dominates and the variance of the union count tends to zero as the CCS finds fewer and fewer of the missed people.

11.2 There are two branches to the further research:

- to extend the design to all counties in England and Wales as well as include Scotland. This requires looking at the issue of outliers produced by the clustering and how to deal with the groups of multiple counties to get the first stage sampling fraction down; and
- to look further at the non perfect dependent CCSs. There is the particular issue of measuring the performance of the estimator in the situations where the bias dominates. As a parallel to the capture recapture work one also needs to introduce DSE into these county level simulations to compare its performance with the regression estimator.

References

Heady, P., Smith, S. and Avery, V. (1994) *National and local demographic estimates*. Census Validation Survey: coverage report, pp 39-44, by OPCS (Social Surveys Division).

Royall, R. M. (1970) *On finite population sampling under certain linear regression models*. Biometrika Vol. 57 pp 377-387.

Royall, R. M. and Cumberland, W. G. (1978) *Variance Estimation in Finite Population Sampling*. JASA Vol. 73 pp 351-361.

Scott, A. J. and Holt, D. (1982) *The Effect of Two-Stage Sampling on Ordinary Least Squares Methods*. JASA Vol. 77 pp 848-854.

Simpson, S., Cossey, R. and Diamond, I. (1997) *1991 population estimates for areas smaller than districts.* Population Trends, 90. (in press)

ANNEX A - CCS Sample Size Calculations

For the super-population model given in Section 4 the BLUP for the total of an age-sex group is the stratum by stratum expansion estimator given by:

$$\hat{T}_{\xi} = \sum_{hd} N_{hd} y_{Shd} \qquad \text{(a1)}$$

where y_{Shd} is the sample mean for the CCS enumeration district count. For this estimator and model the variance of the estimation error is given by:

$$\text{var}_{\xi} = \left(\hat{T}_{\xi} - T\right) = \sum_{hd} \left(N_{hd}^2 / n_{hd}\right)\left(1 - n_{hd}/N_{hd}\right)\sigma_{hd}^2 \qquad \text{(a2)}$$

For a given total sample size of n enumeration districts optimal allocation is used to get the individual stratum population sizes such that:

$$n_{hd} = \frac{n \cdot N_{hd}\sigma_{hd}}{\sum_{g} N_g \sigma_g} \qquad \text{(a3)}$$

For a given population quantity such as the total T with estimator T, one can measure the accuracy of the estimator using the relative standard error (RSE) defined as:

$$\text{RSE}\left(\hat{T}_{\xi}\right) = \frac{\left\{\text{var}_{\xi}\left(\hat{T}_{\xi} - T\right)\right\}^{1/2} \cdot 100}{T} \qquad \text{(a4)}$$

The aim is to design for an RSE of α percent. The variance formula can be written as:

$$\text{var}_{\xi}\left(\hat{T}_{\xi} - T\right) = \sum_{hd} \left(N_{hd}^2 \sigma_{hd}^2 / n_{hd} - N_{hd}\sigma_{hd}^2\right) \qquad \text{(a5)}$$

Now only the first term depends on the sample sizes. Substituting for n_{hd} in terms of n using (b3) in the first term of (b5) gives:

$$\text{var}_{\xi}\left(\hat{T}_{\xi} - T\right) \le \sum_{hd} N_{hd}^2 \sigma_{hd}^2 \cdot \left\{\sum_{g} N_g \sigma_g / n N_{hd} \sigma_{hd}^2\right\} = \left\{\sum_{hd} N_{hd} \sigma_{hd}\right\}^2 / n \qquad \text{(a6)}$$

Using equation (b6) as the variance in the RSE formula gives the approximate sample size required for an RSE of percent as:

$$n = \frac{10^4 \left\{\sum_{hd} N_{hd} \sigma_{hd}\right\}^2}{\alpha^2 T^2} \qquad \text{(a7)}$$

For the actual calculations a design variable is used in place of the Y_i's as these are obviously unknown and the required RSE is 1.0 percent.

ANNEX B - Sensitivity Analysis Plots for Females

Figure B1.

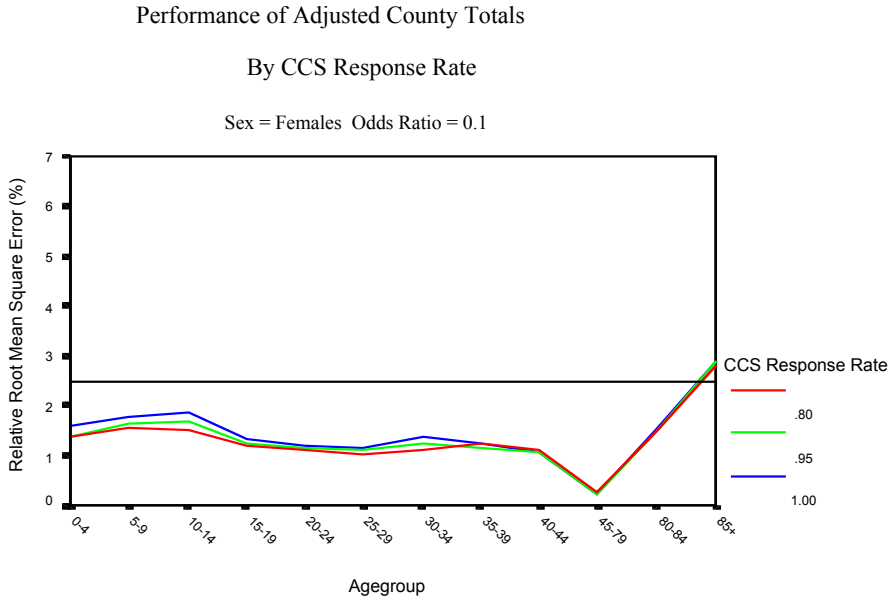


Figure B2.

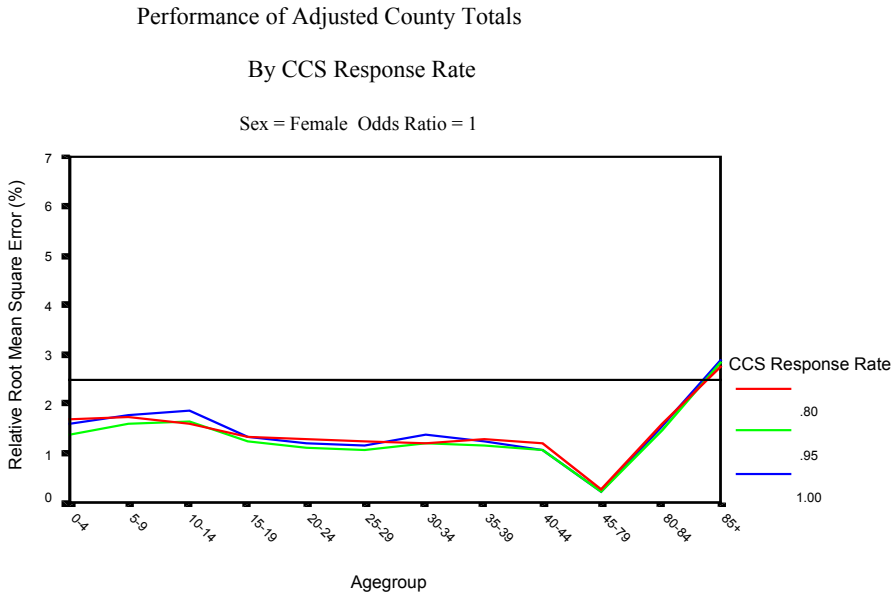


Figure B3.

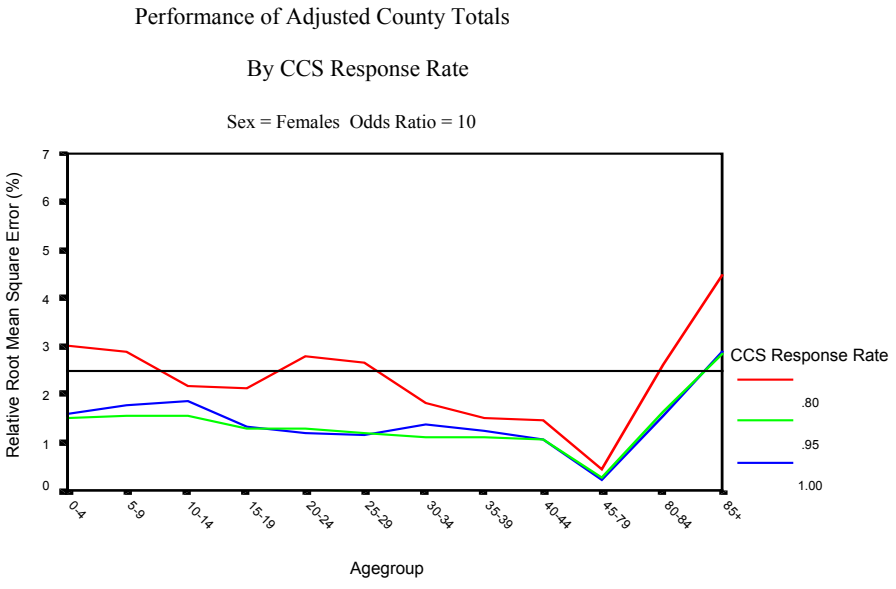


Figure B4.

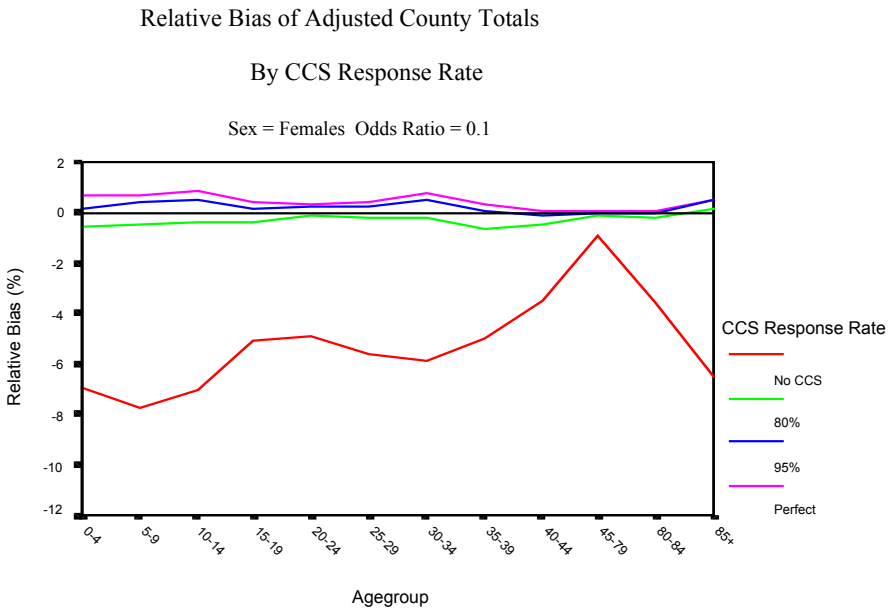


Figure B5.

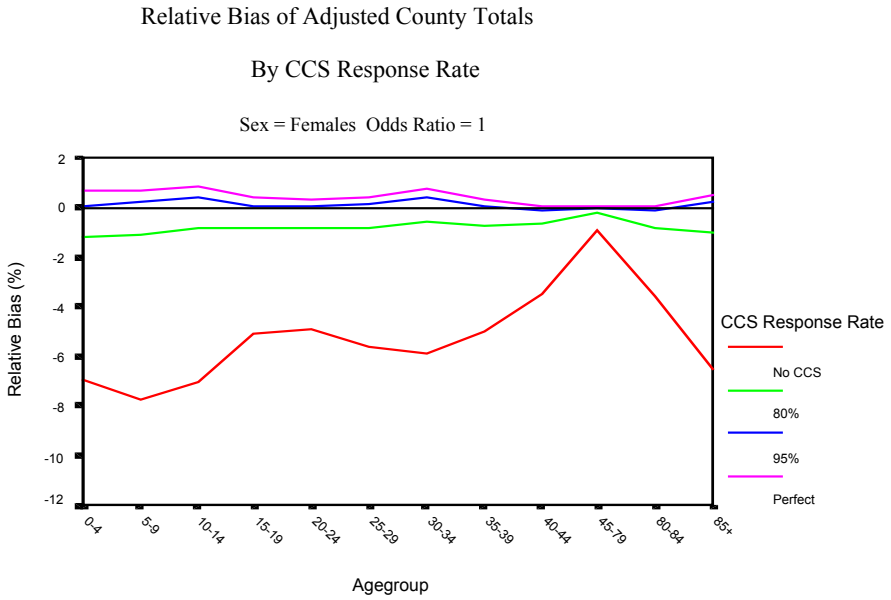


Figure B6.