# 13th Meeting
# of the National Statistics Methodology Advisory Committee

## Overcoverage in the 2011 Census

Owen Abbott (ONS) and James Brown (University of Southampton)

*Executive summary*

For the 2001 UK Census, the Office for National Statistics (ONS) instigated the One Number Census (ONC) project which focused on measuring undercoverage, since overcoverage had historically not been a significant issue in previous censuses. Evidence gathered during the ONC and subsequent studies indicated that overcoverage in the 2001 Census may have been around 0.5%. For the 2011 UK Census it is expected that this figure will be larger due to changes in collection methodology and the rapid change in society and living arrangements. This paper lays out the background to the problem, international practice and developments in this area and the areas of research that will be addressed. Early ideas for measuring are discussed, such as survey and matching based methods. In particular there is an issue of whether a methodology can be developed that integrates overcoverage assessment within the undercoverage assessment framework.

### Aim of paper

ONS is exploring ways for measuring overcoverage alongside the research to develop its methodology for measuring undercoverage. The paper is presented for the opinion of the committee.

*Requested actions from the committee*

The committee is asked to provide any comments at the meeting, particularly to provide their views on the way forward.

## Main issues for discussion

**QUESTION 1:** Do the committee agree that overcoverage, whilst important, will still be secondary in priority to undercoverage?

**QUESTION 2:** Does the committee have any advice on the likelihood of erroneous enumerations in the 2011 Census?

**QUESTION 3:** Are there any other methods for integrating under and overcount within the estimation framework to be used for 2011?

**QUESTION 4:** Does the committee think there are any other sources of data, beyond a PES or matching study, that could help measure overcoverage?

**QUESTION 5:** Which adjustment strategy does the committee favour - Macro level or micro level?

**QUESTION 6:** Does the committee have any advice as to whether ONS should pursue an E-sample type field follow up (given the costs and additional complexity)?

**QUESTION 7:** Does the committee agree with the conclusion that we should adopt a Macro approach, using a single survey combined with matching studies for the measurement of overcount?

# Overcoverage in the 2011 Census

## 1. Introduction

In the UK Censuses, overcoverage has not historically been considered an issue. This is in contrast to the US experience where overcoverage is of a similar magnitude to undercoverage. However, the structure of the population we are trying to count in the UK is changing and the census taking environment is changing to potentially be more similar to that used in the US. Both of these have the potential to increase overcoverage errors in the next census round. This paper presents the historical context before reviewing the measurement of overcoverage in the 2001 Censuses, including a summary of subsequent post-census studies of duplicates. Four categories of overcoverage are defined and the paper then assesses how data from a single fieldwork post-enumeration survey (or a double fieldwork post-enumeration survey as in the US) can be used within the UK estimation framework to integrate overcoverage and undercoverage measurement. The effectiveness of measuring and adjusting for each type of overcoverage is considered, and conclusions are then drawn.

## 2. Overcoverage in previous UK Censuses

### 2.1 1971, 1981, 1991 Censuses

The UK has had a programme of census coverage assessment since the 1971 Censuses. Early post-enumeration surveys (PESs) were on a small scale and looked at global estimates of coverage errors as well as data quality for those that had completed the census questionnaire. The census taking environment was one that controlled carefully the questionnaires in the field. Questionnaires were delivered and collected by an enumerator limiting the possibility of duplicate returns from the same location and giving householders more opportunity to ensure they were counted correctly. In addition, the census counted the population on a present basis (de-facto), although in the 1991 Census information on usual residence was collected. The use of the population present basis ensured that individuals were clear as to the location where they should be counted (and counted once). As a consequence overcoverage in both 1971 and 1981 was estimated to be negligible. The 1981 PES measured overcoverage to be 0.17 per cent (Britton and Birch, 1985).

The 1991 Census was followed by the Census Validation Survey (CVS) which had the twin aims of measuring coverage and quality at the national level. The subsequent analysis (see Heady *et al* (1995) for example) showed that at the national level the undercoverage estimates were deficient while it was assumed that based on information from the CVS overcoverage still remained a second order problem. Most of the overcoverage was a result of the Census imputing for households that were absent around Census night, with subsequent analysis showing that in reality these tended to be smaller than the imputed households. There was evidence of individuals responding at two different addresses, but the number of cases was very small.

### 2.2 The 2001 one-number census

The 2001 Census fieldwork was a major step change from previous censuses. The questionnaires were delivered by enumerators but there was less emphasis on direct contact with householders, and in many cases questionnaires were posted through the letterbox when contact could not be made. Questionnaires could then be posted back and enumerators only followed-up where no questionnaire

**Table 1: Multiple enumerations in the 2001 Census/ LS link**

| Type of multiple enumeration | Number of LS members |
|---|---|
| Students correctly enumerated at term-time and parents addresses | (a) 3,690 |
| Others enumerated more than once | (b) 2,066 |
|     People enumerated in communal establishment and at the family home | 451 |
|     People enumerated at two addresses, one recorded as the address one year ago | 412 |
|     Children of separated or divorced parents who were enumerated as living with both | 262 |
|     People enumerated more than once on the same form | 202 |
|     Duplicate forms: two forms for the same address | 198 |
|     Other cases | 539 |
|     Still under investigation | 2 |
| Total number of LS members enumerated more than once | (a+b) 5,756 |
| LS members enumerated once only | 534,327 |
| Total number of LS members | 540,083 |
| Multiple enumerations (excluding students) as a percentage of all LS members | 0.383 |

was returned. Relaxing the fieldwork controls in this way exposes the census fieldwork to duplicate returns from the same location in a way that had not previously been possible and increases the likelihood of individuals not counting themselves correctly. In addition society was rapidly changing and by 2001 there were many more situations where individuals (adults and children) had multiple locations at which the census could count them than in previous decades. The 2001 Census counted using a 'usual resident' (de-jure) base and again this can cause confusion when individuals have multiple locations that could be their usual residence. Some populations were deliberately double counted. In particular, students were counted at their term-time address, although parents were allowed to include them at their parental address (with a filter question to identify them). This increased the risk of causing overcount if the filter questions were not answered correctly. Overall, these changes in census methodology suggested that overcoverage in 2001 could be more of an issue than in the previous censuses.

The failure of the CVS to be able to adjust the 1991 UK Censuses for use in population estimates led to the one-number census (ONC) project for the 2001 Censuses (see Brown *et al* (1999) and Holt *et al* (2000) for an overview of the approach). A key component of the project was the Census Coverage Survey (CCS); a large-scale area based household survey four weeks after Census Day. The survey fieldwork and estimation strategy both focussed on undercoverage as the priority, as this was still the critical problem to address. However, the CCS was designed to measure overcoverage at the macro-level and therefore if needed make adjustments for it. Results from the CCS put undercoverage at around 6 per cent and overcoverage at around 0.1 per cent (with a high relative error). Based on these results the one-number census database contained no adjustment for overcoverage, but they were fully adjusted to take account of the undercoverage. Steele *et al* (2002) provide a full description of the adjustment methodology.

The estimation process for overcoverage in 2001 involved asking CCS responders where else they could have been counted by the Census. A subsequent matching exercise checked those addresses to

establish whether they had been counted at these alternative locations. This information was used to estimate the levels of overcoverage nationally by age and sex. This revealed that there were increasing issues with children from split families and students were also a problem with them being incorrectly counted at their parents address. However, even within these groups the levels of overcount as measured by the CCS were negligible relative to the undercoverage (Abbott and Brown, 2006a).

However, there were a number of lessons learnt from this approach. Several population subgroups thought to have a high likelihood of overcount fell outside the scope of persons sampled by the CCS or were thought to be unlikely to respond to question four on the CCS. These included persons in communal establishments e.g. young people in armed forces, old people in temporary care, prisoners, and long term hospital patients.

**2.3 Post 2001 Analysis**

Subsequent work by the Office for National Statistics (ONS) used the Longitudinal Study (LS) that links a one percent sample of individuals across censuses. This matching exercise suggested that overcoverage was higher than the CCS had estimated. Full details of the approach can be found in Blackwell *et al* (2003) but the process of matching 1991 individuals to the 2001 database resulted in multiple matches at a relatively higher percentage than the CCS had implied would be the case. Table 1 shows the outcomes of the LS matching and the types of multiple enumerations detected.

The LS found that there were more multiple enumerations in 2001 (5,756 – although 3,690 of these were 'correct' multiple enumerations due to students being counted at both their term-time address and parents address) than in 1981 or 1991 (4,188 and 3,802, respectively, Hattersley and Creeser, 1995, Appendix XIV). However, these figures are not directly comparable. The 2001 Census data included students whom the census intended to be enumerated twice, whereas the 1981 and 1991 Censuses did not. However, the 1981 and 1991 Censuses included cases where individuals could be both a visitor (i.e. a person present who was not usually resident) and an absent resident (they were included as these at their usual residence) whereas the 2001 Census did not. Multiple enumerations in 1981, taking account of records that should not lead to a census overcount (being both a visitor and an absent resident in 1981) represented 0.39 per cent of the population of England and Wales (OPCS, 1988). The comparable figure for 2001 was 0.38 per cent, as shown in table 1.

Further investigation by ONS looked at the existence of duplicates within the final database based on searching for individuals with the same name and date-of-birth (Tromans, 2005). While this exploratory work is not rigorous enough to adjust for overcount, the level of duplicates is indicative of the potential amount of overcount due to this type only. This work confirmed the earlier analysis from the LS both in terms of the level of overcount (the study detected around 0.39 per cent duplicates) and its distribution by age and sex. Figure 1 shows the distribution by age of the duplicates found within the 2001 Census database. It clearly shows that whilst young persons contributed the most duplicates, there is a noticeable proportion of potential overcount of persons across all age groups.
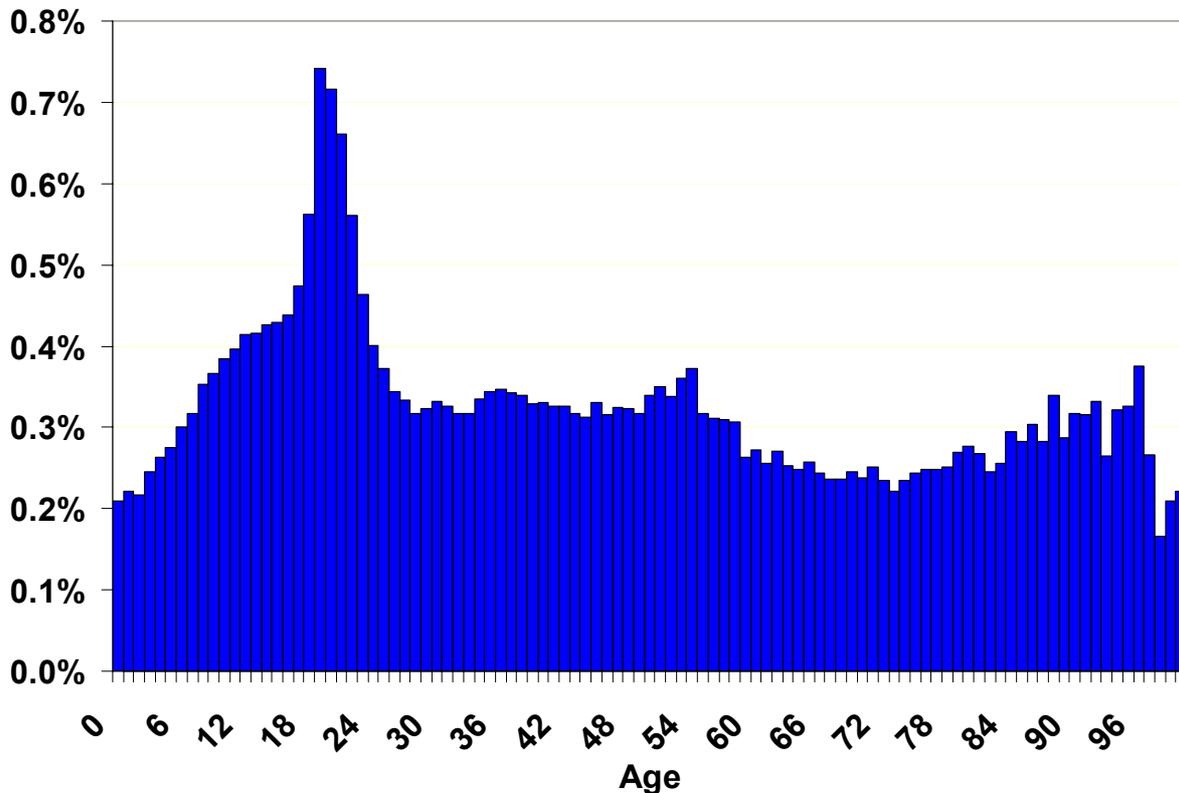
**Figure 1: Potential duplicate persons in the 2001 Census as a proportion of the population estimated by single year of age**

In addition, this study of duplicates within the database identified the likely geographic pattern of the overcount. Only 5 local authorities had duplicate rates of over 1 per cent. Those areas with the most duplicates tended to either have large student populations (places like Manchester and Sheffield), city areas like Central London where individuals may have a second dwelling due to the location of their work and traditional holiday areas where there are a large proportion of holiday homes. The analysis of duplicates did not then go on to try and specify which was the correct location (necessary when actually adjusting for overcount) but one would hypothesise that in areas like Manchester which had the higher levels of duplicates, the duplicates would belong there (students should be enumerated at their term address). In areas like the Lake District and Central London where the levels were lower but nevertheless non-negligible, many of the duplicates would not belong there (holiday homes and second residences used during the week to be near work are not a usual residence according to the 2001 Census definition).

### 3. Overcoverage in the 2011 Censuses

The LS matching exercise revealed in more detail the types of people that were overcounted. This confirmed that it was indeed mainly students, children with split parents and elderly population groups. What these groups have in common is the existence of two locations that could be classified as their 'usual residence'. The concern is that not only are these specific sub-groups of the population on the increase but more generally there are increasing numbers of individuals with second residences (that are not just holiday homes) also potentially increasing the level of overcoverage in 2011. The fieldwork procedures for the 2011 Census are also changing and there is a high probability that post-out with post-back and follow-up will be the model used in many areas, together with the possibility of making a return

over the internet. These changes represent less field control and contact with the public than in 2001. Therefore this approach tends to increase the risk overcoverage with duplicate returns for the same household. This was almost impossible in the past as the enumerator delivery and collection generally prevented multiple returns from the same household. While duplicates can, in principle, be easily removed at processing their existence increase the risk of overcoverage. All this means that when assessing coverage in the 2011 Censuses it will no longer be acceptable to think of overcoverage as an $n^{th}$ order issue relative to undercoverage although we would still expect it to be lower than the levels of undercoverage seen in 2001. The current expectation is that overcoverage in the 2011 Census could be around 1 per cent.

---

**QUESTION 1: Do the committee agree that overcoverage, whilst important, will still be secondary in priority to undercoverage?**

---

### 3.1 Types of overcoverage in 2011

In this paper, we consider that there are essentially four types of overcoverage and these are outlined below.

#### *Type 1 - Duplicate Returns at the same location*
These are where there are two or more returns from *the same* household at *the same* location. This can be either the same individual on a single questionnaire or the same individual on two or more returns (and these returns may not be from the same mode, i.e. paper and internet). These should be removed at processing from the entire database but in reality this will not always happen as two returns may not be identical (slight differences in the number and/or characteristics of the usual residents or one return is for flat one and the second is for flat A).

#### *Type 2 - Erroneous returns*
These are where the householder, enumerator or processing system:
    a) creates fictitious people (e.g. pets);
    b) includes people who are not usual residents of England and Wales (e.g. foreign visitors);
    c) includes a baby born after census day; or;
    d) includes someone who died before census day.
This group are a special problem as the only way you can detect them is by re-visiting householders and asking them to confirm that the people really exist and are usually resident there. This is the achieved via the E-Sample approach used by the US Census Bureau (Hogan, 1993).

---

**QUESTION 2: Does the committee have any advice on the likelihood of erroneous enumerations in the 2011 Census?**

---

#### *Type 3 - Duplicate returns from different locations*
An example of this is where a student is counted at their term-time address (*correct*) and also counted at their home address (*incorrect*) by their parents. This group, if not removed; result in an overcount where they are incorrectly counted. This type of overcount is most associated with students, children of separated parents and people with a second residence. Whilst this group are perhaps easier to detect, the issue of which return is correct requires additional information.

*Type 4 - Counted in the wrong location*

An example is where a student is counted by their parents (*incorrect*), but missed where they should have been counted (their term time address). Nationally, these people are not a problem but once you break counts down by geography they become an overcount in one location and an undercount in the other. This type of overcoverage is difficult to detect and correct.

The aim of the rest of this paper is to consider the possible approaches to estimating overcoverage and how they combat these four types of overcoverage.

## 3.2 An estimation framework for census coverage

The estimation process for the assessment of coverage is fairly similar across many countries. Most use some form of Dual System Estimation (DSE) to estimate those missed by both the Census and PES. What can differ is the way in which this is applied to obtain total population estimates. Brown (2000) provides a summary of DSE methodology, its underpinning assumptions and its history of use in assessing census coverage.

The UK approach to estimation in 2001 was to apply the Dual System Estimator at low levels of aggregation partitioned by age and sex to approximate the homogeneity assumption, which is required for unbiased estimates. In the UK approach local geography is playing the role of the other characteristics used in the US Census Bureau's post-stratified estimator. Both approaches then essentially use ratio estimation to get from the sample to the population. Considering the UK approach, if $cen_i$ is the census count for an age-sex group in small area i (n areas in the sample, N areas in the population), $sur_i$ is the equivalent post-enumeration survey count for the sample areas, $both_i$ are those matched, then the UK approach in 2001 to estimate the population is effectively

$$\sum_{i=1}^{n} \frac{cen_i \times sur_i}{both_i} \Bigg/ \sum_{i=1}^{n} cen_i \times CEN \quad \text{where} \quad CEN = \sum_{i=1}^{N} cen_i \;.$$ A simplified version of the US approach would look

like $$\frac{\sum_{j=1}^{m} w_j cen_j \times \sum_{j=1}^{m} w_j sur_j}{\sum_{j=1}^{m} w_j both_j} \Bigg/ \sum_{j=1}^{m} w_j cen_j \times CEN \quad \text{which simplifies to} \quad \frac{\sum_{j=1}^{m} w_j sur_j}{\sum_{j=1}^{m} w_j both_j} \times CEN \;.$$ The summation

is over m individuals ($cen_j$ is an indicator that takes the value one if individual j is counted by the census in the sample areas) partitioned into post-strata and $w_j$ is the weight resulting from the sampling process and the post-stratification. Both these estimators have the potential for overcount bias and the final US

estimator has the form $$\frac{\sum_{j=1}^{m} w_j sur_j}{\sum_{j=1}^{m} w_j both_j} \times CEN^{adj}$$ where $CEN^{adj}$ has been adjusted down for overcount via the

E-Sample (and other adjustments for imputation done as part of census processing). The full details of the application of the DSE in the US context can be found in Hogan (1993). This is one approach; make a macro adjustment to the census auxiliary so that the ratio estimates account for gross undercount and the macro adjustment accounts for gross overcount. However, this leaves the overcount bias within the DSE and this is therefore not a 'clean' integration.

The alternative would be to make micro level adjustments to the census counts in the survey areas such that the US estimator would look like $\dfrac{\dfrac{\sum\limits_{j=1}^{m} w_j cen_j^{adj} \times \sum\limits_{j=1}^{m} w_j sur_j}{\sum\limits_{j=1}^{m} w_j both_j}}{\Big/ \sum\limits_{j=1}^{m} w_j cen_j \times CEN}$ while the UK

approach would look like $\sum\limits_{i=1}^{n} \dfrac{cen_i^{adj} \times sur_i}{both_i} \Big/ \sum\limits_{i=1}^{n} cen_i \times CEN$ .

This effectively requires the removal of census records that are identified as overcount from the PES areas only, leaving the non-sampled areas 'as is'. The intuitive appeal of this approach, particularly with the UK approach, is that you get the individual DSEs corrected for overcount as well as undercount. However, this then only really gives an estimate of net coverage. While this might be the primary target it is also important to understand the gross levels of the errors as well. As discussed later, this approach also requires a specific type of data collection to make the overcount adjustment, as it must be made at the micro level.

### 3.3 Potential sources of information for assessing overcoverage

The US Census Bureau approach relies on two fieldwork operations, the P-Sample to estimate undercount and the E-Sample to estimate all sources of overcount. The P-Sample is effectively a sample of areas and the survey aims to counts all households and individuals that belong in those areas (as defined by the Census) and is not unlike the 2001 CCS in the UK. On the other hand, the E-Sample is a sample of census returns that are re-interviewed to check that the household and individuals should have been included in the census block where the questionnaire is located. The two samples can be drawn completely independently but there are practical (and cost) advantages to using the same areas for both the P-Sample and the E-Sample and in the US the two samples entirely overlap (Hogan, 1993).

Given the historically low levels of overcoverage in the UK, a second sample specifically for measuring overcount was not considered in 2001 but given the expected increase in this issue, this option is being reconsidered for 2011. From a resource perspective, there is a clear advantage to using a single interview post-enumeration survey to collect information to estimate both overcount and undercount. Therefore, even with increased overcoverage, it is desirable to assess whether a second fieldwork operation is actually necessary in the UK context. It is worth noting that the approach taken by both the Australian Bureau of Statistics (ABS) and Statistics New Zealand uses a single post-enumeration survey to establish where a sample of individuals should have been counted and then how many times and locations where they were counted, including visitors in the sample. This has been successful at detecting reasonable levels of overcoverage (0.8 per cent in Australia in 2006, see ABS (2007) and 0.2 per cent in New Zealand in 2006, see Statistics New Zealand (2007)). A key assumption in the Australian, New Zealand and UK approach is that the post-enumeration survey gets the location of usual residence correct (or does it consistently if two potential locations were sampled for the same person). This assumption has to be considered carefully given the US Census Bureau's experience with ACE in 2000.

Even with a single fieldwork post-enumeration survey, it is not going to be the only potential source of information on overcoverage in 2011. Current research into the 2011 Census questionnaire is investigating collecting more detailed information on second residence and this gives the potential to help find individuals that fall into Type 3 above (counted twice in different locations) as each

questionnaire will potentially identify the other. In situations where you identify individuals that fall into Type 4 (counted once in the wrong location) second residence information might point to the correct location. The issue in this situation is of course finding the overcount in the first place.

---

**QUESTION 3: Does the committee think there are any other sources of data, beyond a PES or matching study, that could help measure overcoverage?**

---

## 4. Strategy options for the 2011 Census

Section 3.2 has highlighted that within the estimation framework that is being used for the 2011 Census, there are essentially two options for integrating overcount measurement with undercount measurement. These two options, which are essentially independent of the sources of information that might be used, are to make either a micro or macro level adjustment.

### 4.1 Micro level adjustment

The principle for this option is to make micro level adjustments to remove overcoverage in the CCS sampled areas, obtaining $cen^{adj}$, such that the resulting Dual System Estimate is an estimate of the true population containing a net adjustment for under and overcount.

Getting $cen^{adj}$ from a single data collection is difficult because by definition if you count them in the post-enumeration survey they should be counted in that location so you tend to learn little about the census overcount within the sampled areas. (The exceptions are those that fall into category one where the matching between the Census and the CCS will potentially find Census duplicates within the sampled areas.)

Thus the critical thing for a micro level adjustment is the ability to gather the information required to decide whether each census observation in the CCS sampled areas is really a usual resident. In this sense it is probably a 'purer' adjustment as the DSE really is aiming to estimate the true population. However, one disadvantage of this approach is that it is not a particularly transparent method of overcount adjustment. Users will not be able to clearly see the under and overcount components of the adjustments.

### 4.2 Macro level adjustment

This option makes macro level adjustments to the auxiliary in the ratio estimator across the whole census. Thus the adjustments for under and overcount are made separately, essentially by applying two ratios to the auxiliary – one for undercount adjustment and the other for overcount adjustment.

Having a single data collection tends to lend itself more to estimating $CEN^{adj}$ (the macro level adjusted census count). In the CCS it is possible to take the same approach as used in 2001 and collect information on where else individuals could have been enumerated and use this information, via matching, to establish adjustment proportions for those individuals in Types 3 and 4. This could also be backed up by matching the census data (either all of it or a sample) to detect duplicates (Type 3s).

The advantage of using the $CEN^{adj}$ approach is that the estimation strata used for overcount do not need to overlap directly with those used for undercount. In terms of processing, it is also possible to produce initial estimates on a sequential basis adjusting for undercount (still expected to be the main issue in

2011) and then make the overcount adjustments at the end once processing is completed as matching across areas is only possible once all areas have been processed.

This allows any Quality Assurance process (see section 4.4) the ability to identify early those areas where the risk of failure of the undercount measurement process is high as well as highlight any areas where we might expect the subsequent overcount adjustments to have an impact. In addition, the revision of any such cases is also simpler as the adjustments are relatively easily undone or altered.

Lastly, the method is clearly transparent to users in that separate adjustments are made. This will have clear benefits when consulting with users about the methodology and also when explaining the results, especially to non-technical audiences.

---

**QUESTION 4: Are there any other methods for integrating under and overcount within the estimation framework to be used for 2011?**

---

**QUESTION 5: Which adjustment strategy does the committee favour - Macro or micro level?**

---

### 4.3 Measuring all types of overcount

This section outlines how each type of overcount can be measured, bearing in mind the two options for adjustment discussed above.

### 4.3.1 Type 1 – duplicates at the same address

This type of overcount can be dealt with at the processing stage through a relatively simple automated matching and decision algorithm, prior to any coverage assessment. Thus this can be applied whether adopting micro or macro overcount adjustments (and therefore is not an issue for either). This is the approach used in 2001, and it is expected that the system used can be improved to eliminate the few cases that were not identified and resolved.

The 2011 Census will use an address frame that should enable questionnaires (both paper and internet) issued and returned from that address to be linked automatically, thus providing very tight matching criteria for each address. In addition, the process could be widened to match not just those records for the same address but also those close by (perhaps within the same postcode). This would cover some of the duplicate cases that might arise from mis-delivered questionnaires or localised geo-coding problems, although care would have to be taken.

Whilst this process requires some further development work this is not expected to be a big issue beyond the algorithm for merging the data where different data exists (i.e. what to do if you make a match but some of the variables are different).

### 4.3.2 Type 2 – Erroneous Enumerations

A single fieldwork process cannot identify and estimate for this type of overcount. The post-enumeration survey should not count them where the Census counted them (as they do not exist or are not really a usual resident on Census day), and they will not be identified elsewhere by the post-enumeration survey as they do not have a usual residence elsewhere.

The only way to measure this type of overcount is to have a second field process to effectively resolve those census records that are not found in the CCS (because as mentioned above we expect the CCS to not enumerate them as it is a higher quality counting process). The US E-sample is basically this second field process.

The question is then will this Type 2 overcount be important in the UK in 2011 and thus how much protection do we want to minimise the risk of this being a significant issue. Part of the issue in the US relates to the way the Census is conducted. It allows enumerators to collect proxy returns (which can be very inaccurate) and at the processing stage large-scale imputation is done for households where the enumerator states a household exists but for which no (or very limited) information could be collected. This is important as the US Census database is not adjusted for undercoverage and these processes help make sure all individuals are included. However, in the UK in 2001, such information was not included on the Census database as the one-number census process adjusted the final database for undercoverage so there is less exposure to this Type 2 overcount.

However, this type of overcoverage does not just occur due to field and processing procedures. There is evidence in the US Census of enumerators 'curb-stoning' where they simply make-up a return for a household on their address list. In the UK there is no evidence that this has been an issue but in the past there has never been an address list so enumerators have not had the same pressure to obtain returns from specific addresses. However, the intention in 2011 is to create an accurate address list as a basis for the fieldwork. This may create an environment in which enumerators do feel pressured to get returns from a specific location (especially if pay is linked very closely to achieving a return) and result in 'curb-stoning'.

If 'curb-stoning' does take place, the single fieldwork post-enumeration approach will over-estimate the population as these erroneous census returns inflate the DSE. The ONS is currently considering a survey to measure quality after the 2011 Census, which will involve re-interviewing households that provided a census return. Whilst the sample size of the quality survey is unlikely to allow estimation at low levels of geography, it would reveal if there has been an erroneous enumeration issue at a national level. This would provide evidence to make an additional adjustment if deemed necessary, although this is easier to integrate if we adopt a macro adjustment approach to overcount adjustment.

In addition to this, the single fieldwork post-enumeration approach may give an indication of a problem as these erroneous census returns will not match to the survey in the sample areas. 'Curb-stoning' will result in a survey return from the same location having no individuals in common with the census return for the same location. Whilst this may just reflect a poor within household count by both the census and the survey it could also indicate an erroneous census return. Further investigation of significant cases (or a sample of all unmatched records) using additional management information might highlight where there has been an issue. This is the type of evidence that in combination with a quality survey would justify an additional adjustment.

The US Census Bureau uses its E-Sample to estimate $CEN^{adj}$ although given the overlapping samples it would also be possible to estimate $cen^{adj}$. Thus if ONS consider that the risk of erroneous enumerations is great and a micro adjustment approach is preferred, then the consideration of running a similar style E-sample approach is essential. This could be achieved by boosting the sample size of the quality survey.

> **QUESTION 6: Does the committee have any advice as to whether ONS should pursue an E-sample type field follow up (given the costs and additional complexity)?**

### 4.3.3 Type 3 – Duplicate returns from different locations

One of the proposed new topics for inclusion in the 2011 Census is a question on second residence. This would ask where else in the UK each resident might be counted. This information could potentially help estimate global adjustments for those in Type 3 via large-scale matching between census questionnaires. This would be possible once the database had been processed and initial work suggests the costs may not be prohibitive, but there are processing timetable implications.

In addition, it would be difficult to identify from the Census alone which of the two returns was the correct location. This is not an issue if the two returns are in the same estimation stratum (age-sex group by hard-to-count area in the UK) as you simply need to adjust down the Census auxiliary for the double-count. However, when the two returns are more distant (say London and Southampton) we would need to identify the correct one. This is where the CCS can help combined with the Census.

In 2001 we have noted that the CCS alone was poor at identifying alternative locations. Section 4.3.4 discusses how better fieldwork procedures can improve this, but the additional information in the Census can also help. The CCS may fail to collect information on alternative locations but matching to the Census may identify a second residence. This can then be searched as it would be if identified directly by the CCS giving us a second method of identifying alternative addresses for the CCS respondents.

However, even without the second residence data, we can still match the census data against itself to find duplicates. This could use a number of searching strategies to efficiently detect duplicates. We could use our knowledge of populations at risk of overcoverage to draw samples of those populations to carry out searches, and we can also simply look for name duplicates after blocking by date of birth. These strategies need further development work, but both can be used within either the micro or macro adjustment approach.

### 4.3.4 Type 4 – Counted in the wrong location

The problem with having a single data collection is that in 2001 it was not sufficiently robust to estimate the extent of Type 3 and 4 overcount. However, the CCS fieldwork procedures and questionnaire design focused almost entirely on undercount with the overcount question just being added on the end without much testing and development work to check the effectiveness of the question. We need to ensure, like the ABS, that our questionnaire not only finds people where they should be but also robustly identifies where else they could be. This may require targeted questions for individuals with characteristics in the CCS that make them susceptible to overcount (for instance children in single parent families and students).

In order to measure these type 4 errors, as well as ensuring the CCS counts residents correctly it also needs to ensure it counts the visitors on census day and where those visitors could have been counted. This is to ensure that when matching we can use this information to detect, at micro level, census residents who should have been visitors. This can be used to make adjustments at a micro or macro level.

**4.4 Quality assurance of overcoverage post 2011**

It is usual practice to carry out comparisons between the census population estimates and relevant population level data, such as demographic mid-year population estimates, sex ratios and administrative sources. This Quality Assurance process was, in the 2001 Census, focused on undercount as an issue. White *et al* (2006) provide a summary of the strategy and sources used.

The LS that came later did detect more overcount than was previously thought. The question is whether a similar approach to the LS could be used at the end of processing as a QA. For example, post-analysis by ONS has looked for duplicates based on name and date-of-birth. While this is not entirely accurate it is a relatively efficient way of getting indicators of the level of overcount that *more sophisticated* estimates from the CCS can be compared to. In addition, the LS linkage exercise may be able to be completed earlier in 2011 which will provide the opportunity to use this within the QA process.

**5. Conclusions**

It is clear that overcoverage will be a more important issue for coverage assessment in the 2011 UK Census than it was in 2001. Societal change, and the method by which the 2011 Census will be executed will undoubtedly lead to increased levels of overcoverage.

As a consequence, the coverage assessment strategy used in 2001 needs to be enhanced. Current thinking is focused around a single large-scale post-enumeration survey that is supported by information from a much smaller scale quality survey and various matching exercises which look for duplicates. The sample size and scope of any quality survey has yet to be decided and its use to assess overcoverage potentially implies a survey on a larger scale than might otherwise have been the case. This increase in size is closely related to the likelihood of a significant number of erroneous enumerations. If there is an indication that this will be a big issue, then the survey needs to effectively become the equivalent of the US E-sample. The cost of such an approach is not insignificant.

Both the single and dual survey approach can provide an adjustment for overcoverage that can be relatively easily integrated into an estimation framework similar to that used in 2001. The key choice when integrating the overcount and undercount measurement within this framework is between a micro level adjustment or a macro level adjustment. The micro level adjustment is a 'truer' adjustment, but it is technically more difficult and would require more information (as to the choice of which records to remove). The macro level approach has the advantages of transparency and timeliness but it does leave some risk of overestimation in the DSE. On balance, due to its simplicity, the current assumption is that a macro level approach should be adopted. Further work should be undertaken to explore the effectiveness of this approach.

This paper has concentrated on the estimation of overcoverage at the level for which direct estimates of the population are produced. In the UK in 2001 this was for groups of Local Authorities with populations of around 0.5 million. Small area estimation techniques were used to produce the full set of Local Authority age-sex population estimates and this paper has not addressed how overcoverage would be integrated into these estimates. In addition, the one-number census in 2001 used coverage models to create weights that were used in the imputation that created the final database adjusted for undercoverage (see Steele *et al*, 2002). This paper has not addressed how overcoverage would be reflected in a coverage adjusted census database in 2011, although early research has begun to think about the issues involved. What is clear is that the possibility of marking records to give them a weight of

zero (effectively deleting the record from the census database) for outputs has not been ruled out. These are issues that need to be fully addressed once there is an agreed strategy on the estimation of overcoverage which is what this paper has addressed.

---

**QUESTION 7: Does the committee agree with the conclusion that we should adopt a Macro approach, using a single survey combined with matching studies for the measurement of overcount?**

---

**References**

Abbott, O. and Brown, J. (2006a) A review of the 2001 One Number Census methodology and lessons learnt. Paper presented at GSS Methodology Conference, London, June 2006. Available at www.statistics.gov.uk/events/gss2006/downloads/D1Abbott.doc

Abbott, O. and Brown, J. (2006b) Coverage assessment and adjustment in the 2011 Census, 2006 Proceedings of the American Statistical Association, Survey Research Section [CD-ROM], American Statistical Association, Alexandria, VA.

ABS (2007) Census of population and housing – details of undercount. Cat no. 2940.0. Australian Bureau of Statistics, Sydney. Available at http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/2940.0Aug%202006?OpenDocument

Blackwell, L., Lynch, K., Smith, J. and Goldblatt, P. (2003) Longitudinal Study 1971-2001: Completeness of Census Linkage. Series LS No.10. London: Office for National Statistics. Available at www.statistics.gov.uk/downloads/theme_population/LS_no10.pdf

Britton, M. and Birch, F. (1985) 1981 Census Post-Enumeration Survey. An enquiry into the coverage and quality of the 1981 Census in England and Wales. London: HMSO. Available at www.statistics.gov.uk/ about/data/methodology/specific/population/LS/resources/pes.asp

Brown, J. J., Diamond, I. D., Chambers, R. L., Buckner, L. J., and Teague, A. D. (1999) A methodological strategy for a one-number census in the UK. *J. R. Statist. Soc. A*, **162**, 247-267.

Brown, J. J. (2000) Design of a census coverage survey and its use in the estimation and adjustment of census underenumeration. PhD Thesis. University of Southampton, Southampton.

Hattersley, L. and Creeser, R. (1995) Longitudinal Study 1971-1991: History, organization and quality of data, London: HMSO. Available at www.statistics.gov.uk/downloads/theme_population/LSno7_v4.pdf

Heady, P., Smith, S. and Avery, V. (1994) 1991 Census Validation Survey: coverage report, London: HMSO. Available at www.statistics.gov.uk/ about/data/methodology/specific/population/LS/resources/cvs.asp

Hogan, H. (1993) The 1990 Post-enumeration survey: Operations and results. *Journal of the American Statistical Association*, **88**, 1047-1060.

Holt, T., Diamond, I. D., and Cruddas, M. (2001) Risk in official statistics: a case-study of the 2001 one-number census project. *J. R. Statist. Soc. D*, **50**, 441-456.

OPCS (1988) Census 1971-1981 The Longitudinal Study, London: HMSO

ONS (2005) One Number Census Evaluation Report. Available at
www.statistics.gov.uk/census2001/pdfs/onc_evr_rep.pdf

ONS (2006) 2011 UK Census Coverage Assessment and Adjustment Strategy, Advisory Group paper
(06)16. Available at www.statistics.gov.uk/ census/pdfs/ag0616.pdf

Statistics New Zealand (2007) A report on the 2006 Post-enumeration Survey, Statistics New Zealand,
Wellington.

Steele, F., Brown, J. and Chambers, R. (2002) A controlled donor imputation system for a one-number
census. J. R. Statist. Soc. A, 165, 495-522.

Tromans, N. (2005) Discussion paper for assessing the level of overenumeration in the Census.
Unpublished paper. Available on request.

White, N., Abbott, O., and Compton, G. (2006) Demographic analysis in the UK Census: a look back to
2001 and looking forward to 2011. 2006 Proceedings of the American Statistical Association, Survey
Research Section [CD-ROM], American Statistical Association, Alexandria, VA