2021 Census Coverage Survey Design Strategy – V4.0

**Project**: Census Statistical and Outputs Design (CSOD)

**Project SRO**: Iain Bell

**Project Executive**: Shaun Garvey

**Document Control**

|  |  |
| --- | --- |
| **Product Reference** | To be assigned by the Census Programme Management Office when final version is available |
| **Description** | 2021 Census Coverage Survey (CCS) design strategy, developed by Methodology for the 2021 Census |
| **Owner** | Adriana Castaldo |
| **Approver** | Owen Abbott |
| **Consulted** | * Census Statistical and Outputs Design (Cal Ghee/Orlaith Fraser) * CCS Field Operations and CCS Working Group (Tansy Arthur) * Census Transformation Programme Research Assurance Group (CRAG) |
| **Informed** | * Census Design Forum |
| **Related Products** | * 2021 CCS Requirements – v5.1 (Adriana Castaldo) * 2021 Census Coverage Assessment Strategy (Viktor Racinskij) * 2021 Census Coverage Adjustment Strategy (Alison Whitworth) * Hard to Count index for the 2021 Census (Ercilia Dini) |

**Version Control**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Version** | **Created By** | **Date** | **Status** | **Changes Made** | **Made By** |
| 1.0 | Adriana Castaldo | 28 June 2018 | 1st Draft | With input from Owen Abbott, Paul Smith (University of Southampton), Ercilia Dini and Viktor Racinskij | Adriana Castaldo |
| 2.0 | Adriana Castaldo | 7 August 2018 | 2nd Draft | For submission to CRAG, sponsored by Salah Merad | Adriana Castaldo |
| 3.0 | Adriana Castaldo | 12 September 2018 | 3rd Draft | Revised based on comments from CRAG, CSOD, CCS Field Operations and CCS Working Group | Adriana Castaldo |
| 4.0 | Adriana Castaldo | 14 September 2018 | Final Version | Signed off by Owen Abbott | Adriana Castaldo |

# Executive Summary

The Census Coverage Survey (CCS) is a large household survey conducted shortly after the census, with the aim of measuring and adjusting for under- and over-count. In 2011 a CCS was conducted in England and Wales six weeks after census day. The CCS sample included 17,400 postcodes (or 1.5% of all postcodes) containing nearly 340,000 households. The survey used a stratified two-stage cluster sample, first selecting output areas by Local Authority and the Hard to Count index, and then selecting approximately half the postcodes from within those output areas. The 2011 CCS sample design was an improved version of the successful 2001 CCS.

In preparation for the 2021 Census, ONS has been reviewing the 2011 CCS design and has been looking at ways to improve the design of the survey for 2021. This paper provides an overview of the research that has been conducted to date, and provides a list of recommendations for the design to be adopted in 2021. The research suggests using a similar sample design to that adopted in 2011, but considering using a reduced sampling fraction of 25% rather than 50% at the second stage of sampling.

The same stratification variables as in 2011 are recommended, that is, Local Authority and an updated version of the Hard to Count Index. The updated version of the Hard to Count index to be used for the stratification of the CCS sample will be the same index that is being developed for the planning of the 2021 Census field follow-up. However, afinal decision on whether we will be using both the ‘willingness’ and the ‘digital’ domain of the 2021 index, or only the ‘willingness’ domain, has not been made yet and requires further research.

Further analysis is also recommended to assess the option of using a sample allocation method for the 2021 CCS, that is less reliant on information about the 2011 Census coverage patterns.

Contents

[Executive Summary 2](#_Toc524711678)

[1. Summary of the 2011 CCS Design 4](#_Toc524711679)

[2. 2021 CCS Design: overview of work done so far 5](#_Toc524711680)

[3. Choice of Final Sampling Unit 6](#_Toc524711681)

[4. Choice of: One or Two-stage Design, PSU, and Sampling Fractions 8](#_Toc524711682)

[5. Stratification and Hard to Count Index 9](#_Toc524711683)

[6. Sample Size and Allocation 11](#_Toc524711684)

[7. Other Sample Design Questions 13](#_Toc524711685)

[7.1. CCS Sampling Strategy for Communal Establishments 13](#_Toc524711686)

[7.2. ‘Flexible Sample’ Option 14](#_Toc524711687)

[8. Summary of Recommendations for the 2021 Design 15](#_Toc524711688)

[References 16](#_Toc524711689)

# 1. Summary of the 2011 CCS Design

The 2011 Census Coverage Survey (CCS) was a voluntary, interviewer-led doorstep survey of about 1% of households (or 1.5% of postcodes) in England and Wales. The survey was carried out independently of the census and took place six weeks after census day (from 9 May 2011 to 2 June 2011). The main objective of the CCS was to measure coverage in the 2011 Census and adjust the census database for those estimated to have been missed (ONS, 2012a, 2012b). The 2011 CCS sample design was an improved version of the successful 2001 CCS (Brown, Abbott and Smith, 2011).

In 2011 the CCS used a stratified two-stage cluster design (ONS, 2010), with stratification by Local Authority (LA) and the National Hard to Count index (Hopper, 2011) - which had five levels, ranging from 1 (least hard-to-count) to 5 (hardest-to-count). In the first stage, Output Areas (OAs) were sampled as Primary Sampling Units (PSUs) within each stratum. In the second stage, postcodes were sampled as Secondary Sampling Units (SSUs) within the selected PSUs.

The PSU sample rate was approximately 3% of OAs per LA on average, while the SSU sample rate was 50% of postcodes per OA. One key change from the 2001 sample design was that the sample was less clustered. In 2001, 5 postcodes per OA were selected. In 2011 the original plan was to reduce this to 3 and allow more OAs to be selected to maintain the same overall sample size (Brown, Abbott and Smith, 2011). However, due to the varying number of postcodes contained in OAs, this was later changed to sample half the postcodes in each OA. For instance, in OAs with many small postcodes, selecting only three postcodes did not give a large enough sample of households to produce reliable Dual-system estimates at the OA level (Hill and Abbott, 2010).

The overall sample size was circa 17,400 postcodes covering nearly 340,000 households; however, 16,700 postcodes were used in the estimation process (ONS, 2012b, p. 5). A design variable based on the 2001 household imputation counts was used to determine the sample size of OAs within each stratum, with a constraint of a minimum of 1 OA per stratum and a maximum of 60 OAs in an LA. This design variable skewed the sample towards those strata where response was expected to be low or extremely variable, which tended to be urban and inner-city local authorities where response was low in the 2001 Census.

Simple random sampling (SRS) was used in both stages of sampling. All the households in the selected postcodes were firstly independently listed by field staff using postcode maps, and then they attempted to carry out short doorstep interviews with all listed households.

Other features of the 2011 CCS included some flexibility in the sample design strategy. In some areas where the census had a lower response rate than expected, additional postcodes were sampled for the CCS so that a better estimate of those that were missed could be obtained.

Results from the evaluation of the 2011 CCS sample design can be found in ONS (2013, pp. 14-15). Successes and lessons included:

* The use of the LA stratification, which allowed users to see how the sample size was computed directly across all LAs rather than it being a function of a random selection process was a success
* The Hard to Count (HtC) index, which helped reduce the variability in census response rates and the quality of the estimates, through an improved census fieldwork and CCS design was successful. Improving the HtC index is a very important aspect of the census and the CCS design, particularly considering how changeable coverage patterns between censuses can be. The refinement of the HtC index for the 2021 Census is being researched as part of a separate work package (Dini, 2018)
* An adjustment was successfully introduced during processing where the 2011 CCS sample was detected as not being representative (or balanced) with respect to census non-response (ONS, 2012c). This recognised that it is possible to draw an extreme sample by chance, even if the sample is well designed. The methodology detected such cases and provided a method to adjust the resulting population estimates
* The information used from the Postcode Address File (PAF) underestimated the number of households in the postcodes, leading to more households that needed to be interviewed than planned, and thus higher workloads for interviewers. This risk was expected and it was mitigated to a certain extent by introducing flexible working of additional hours for the interviewers. However, this lesson highlighted the need for improved estimates of the number of households within the sampled areas
* Boosting of the CCS sample in areas where census response was much lower than expected was found difficult to implement, which limited the size of the boost. The indications were that it made little difference to the variability of the estimates due to its small nature, and given that the areas that were boosted tended to have large sample sizes without the boost. It was recommended that boosting be considered much earlier in the planning process and consideration given to only use it if it can be shown to significantly improve the estimates.

# 2. 2021 CCS Design: overview of work done so far

In preparation for the 2021 Census, ONS evaluated different options for the CCS design, using the 2011 design as the baseline approach, and looking at ways to achieve more precise census estimates while keeping costs within acceptable levels (Castaldo and Abbott, 2016).

The initial focus was on reviewing the high-level sample design, and exploring options that essentially affected the level of clustering. This included whether to adopt a one- or two-stage cluster design, the PSU/SSU sampling rates, and the choice of sampling units between OAs and Lower-Layer Super Output Areas (LSOAs) for the PSUs, and between postcodes or addresses for the SSUs (i.e., the final sampling units). These questions were investigated using simulation studies, with an assumption that the following key features of the CCS would be consistent with 2011:

* Overall sample size
* Stratification by LA and a revised version of the national Hard to Count index
* Fieldwork mode being interviewer-led doorstep
* Short questionnaire and usual residence collection base
* 6 weeks after census
* No overlap between the CCS and the census field-work

The results from these simulation studies are described in more detail in section 3 and 4 below.

After the analysis assessing different design options that affected the level of clustering, the questions of what stratification variables, sample size and sample allocation method to be used for the 2021 CCS were addressed. Thus, stratification and the Hard to Count index are covered in section 5 of this paper, while sample size and allocation are covered in section 6. The discussion in section 5 and 6 is not based on new simulation studies.

The Hard to Count index for the 2021 Census is being developed under a separate workstream (Dini, 2018) for both the planning of the census field follow-up (among other things) and for use as a stratification variable in the CCS sample. Thus, section 5 draws from the research described in Dini (2018).

Extensive simulation studies to investigate different options of sample size and sample allocation method were conducted during the design phase of the 2011 CCS (Abbott, 2008b), thus section 6 mainly draws from the results of that analysis. New simulation studies are being conducted to assess the effect of using different sample allocation methods on the precision of the census estimates using the estimation method that is being developed for the 2021 Census. The results from this additional analysis will be available in subsequent papers.

Some additional aspects of the sample design for the 2021 CCS including the sample design strategy for communal establishments and the use of the flexible sample option are discussed in section 7.

Each of the following sections, from 3 to 7, include appropriate recommendations, thus the paper concludes with a summary of all recommendations drawn from the work on the 2021 CCS design completed to date.

# 3. Choice of Final Sampling Unit

The first question that was investigated as part of the CCS design evaluation was whether for the 2021 CCS we could move from an area based to an address based sample design (i.e., whether we would be sampling addresses rather than postcodes as the final sampling units). This important question was addressed at the beginning of the project, given that it would be a fundamental change in sampling approach.

The main reason for considering the use of addresses as opposed to postcodes as the final sampling units for the CCS in 2021 was the recent development of the AddressBase product (Ordnance Survey, 2014). The 2011 Census Address List demonstrated that it is possible to derive a high-quality frame through combining information from the Postcode Address File (PAF) with Local Land and Property Gazetteers (LLPGs), and the lessons learnt from that exercise are being embedded in the development of AddressBase. This suggested that the likely quality of AddressBase should be high and therefore we could consider using it as a frame for the CCS, as well as for the census. In the meantime, while carrying out the current research, quality assurance of AddressBase was being undertaken by the Address Register team at ONS to provide evidence to assess that assumption.

Thus, as a first step, the advantages and disadvantages of moving from an area based to an address based sample design for the CCS were discussed by Castaldo and Abbott (2016). The paper also included a work plan setting out what evidence was necessary (and achievable as part of this project) to inform the decision to move to an address based design. The research plan was divided in three phases:

1. Analysis of the impact of moving to an address based design on the statistical precision of the census estimates
2. Analysis of the impact on the costs of running the CCS (i.e., costs of interviewing and travel costs)
3. Consideration of the practicalities of implementing the change, and of other processes that would be affected by such a change (e.g., which estimation approach should we use were we to move to an address based sample design).

The evidence from the research in phases 1 and 3, along with the proposed approach to the analysis of the costs (phase 2), was presented and discussed by Castaldo (2017).

The key findings presented in that paper, were as follows:

* The estimated ‘gains’ in precisions of the census estimates, associated with moving from a postcode to an address based sample design, were not as big as expected (for more details on this, see section 4 below).
* Moving to an address based sample design meant too much reliance on the quality of AddressBase and a higher risk that the census and CCS would not be independent.
* There were no available measures of the quality of AddressBase yet, thus making it difficult for us to rely on its quality at that point in time.
* Due to the Census and CCS no longer being independent, two different dual system estimates (DSEs) would need to be constructed: one to account for missed people in addresses that are on the frame, and one to account for missed people in addresses that are not on the frame. This added complication into the estimation strategy.
* Lastly, from a practical point of view moving to an address based design would mean a complete review of the CCS design and operations. This implied potentially many other sources of risks and costs (along with possibly some savings) that were difficult to measure in advance - given that this would be a completely new way of carrying out the CCS.

Based on the assessment done in phases 1 and 3 described above, it was agreed not to pursue the option of an address based sample design for the 2021 CCS. However, it was also agreed that the work started under phase 2 would keep including the option of an address based design, primarily for research purposes - due to the possible wider applications of the research. For example, although using an address based sample design may not be a viable option for the CCS due to the higher risk of the CCS not being independent of the census, it could be a good option for other sample surveys, for which independence of the census is not required. Thus, the decision for the 2021 Census Coverage Survey design was to retain the sampling unit as unit postcodes.

# 4. Choice of: One or Two-stage Design, PSU, and Sampling Fractions

The question of the number of stages, the choice of PSU (were we to use a two-stage design), and the choice of PSU/SSU sampling fractions, were investigated in Castaldo and Nikolakis (2018). That paper investigated the effect of different design options for the 2021 CCS on the statistical precision of the census population estimates. The aim was to measure the gain in statistical efficiency that could be achieved through the adoption of alternative sampling approaches, assuming the same overall sample size, and using an estimation method similar to the 2011 method.

The analysis used simulation studies and was carried out for three estimation areas:[[1]](#footnote-1) Manchester (chosen as a predominantly urban area), South West Wales and Powys (chosen as a predominantly rural area), and South-East Hampshire (chosen as a mix of urban and rural). Six different types of sample designs were considered (all with proportional sample allocation), as described in Table 4.1 below.

**Table 4.1: Sample design options considered for the 2021 CCS**

|  |  |  |
| --- | --- | --- |
| **PSU** | **SSU** | **Description of option** |
| OA | postcode | 2-stage clustered sample with postcodes as SSUs, i.e., as was done in 2011 |
| OA | address | 2-stage clustered sample with addresses as SSUs, so less clustered than in 2011 |
| LSOA | postcode | 2-stage clustered sample with postcodes as SSUs but LSOAs as PSUs |
| LSOA | address | 2-stage clustered sample with addresses as SSUs and LSOAs as PSUs |
| - | postcode | 1-stage clustered sample with postcodes as sampling unit |
| - | address | 1-stage unclustered sample of addresses (or households) |

For each of the three estimation areas and each of the six sample design options, the relative bias (RB) and the relative standard error (RSE) of population estimates were computed for a range of PSU/SSU sampling fractions. For example, for Manchester, we took the actual CCS sample size and fraction that were achieved in 2011; thus, the first option that we tested was for a PSU sample rate of 4% and SSU sample rate of 50%. Starting from this first option we then computed the resulting number of sampled households (which was approximately 4,600), and looked at the effect of decreasing the SSU sample fraction from 50% to 25%, 20%, 15%, 10% and 5%, while increasing the PSU sample fraction accordingly (that is, so to keep the overall number of sampled households approximately equal to 4,600).

For each of the three estimation areas, each of the six sample design options, and each combination of PSU/SSU sampling fraction, 400 replicates of census realisations and corresponding CCS responses were simulated. We assumed that the generated CCS had perfect response (and therefore counted the population perfectly within its sample, so no need for DSEs), and a ratio estimator was used to obtain population estimates by age and sex for the Estimation area. Another key assumption was that the sampling frame had perfect coverage, that is, covered 100% of addresses. As the true population sizes were known, the relative bias and relative standard errors were calculated across all of the simulations to provide measures of bias and precision.

The relative bias was negligible for all the different sample design options and for all combinations of PSU and SSU sampling fractions for the three estimation areas, as would be expected by design from an approximately unbiased estimator.

The results on the variability of the population estimates (as measured by the RSE) showed that in Manchester moving from the 2011 design to an address based SRS (i.e., a fully unclustered sample) could reduce the RSE by approximately 24% (given the sample allocation is proportional). However, the scale of reductions was less in the other two areas tested, where the respective gains were approximately 8% for South West Wales and Powys, and 12% for South East Hampshire (Castaldo and Nikolakis, 2018). The highest gains in statistical precision were observed for the smallest SSU sampling fractions. This shows how the precision changes as the level of clustering decreases; thus, the question is whether that gain in precision can be achieved with little change in costs.

The option of using LSOAs as PSUs gave the worst results in terms of the variability of the estimates across all scenarios, apart from the cases of very small SSU sampling fractions and when using addresses as SSUs. Thus, the results in that paper suggested that using OAs as PSUs is a better option than using LSOAs.

**Thus, the main conclusions from this analysis were as follows:**

1. **A two-stage design like that adopted in 2011 is preferable to a one-stage design**
2. **OAs as the PSUs should be retained**
3. **The SSU sampling fraction to be adopted in the 2021 CCS sample could be reduced from its 2011 level of 50% postcodes per sampled OA, to 25% postcodes per sampled OA. This would reduce the level of clustering in the sample (thus improving statistical efficiency) with possibly only a slight increase in costs.**

The indications from the analysis of the costs of the different sample design options (Castaldo and Harfoot, 2017, Castaldo, Cockings and Harfoot, 2018, Cockings, Harfoot and Martin, 2017) are that moving to the proposed design does not increase costs significantly, thus the proposed design seems a good balance between cost and precision. Furthermore, a smaller-scale simulation study is recommended to assess the impact of reducing the SSU sampling fraction from 50% to 25% on the census estimates, using the improved estimation method that is under development for the 2021 Census.

# 5. Stratification and Hard to Count Index

As mentioned in section 1 of this paper, in 2011 the CCS sample was stratified by LA and the National Hard to Count Index (Hopper, 2011). Stratification by the Hard to Count (HtC) index is an important aspect of the CCS design, as the objective of the CCS is to measure census non-response, thus, the survey must cover both areas that are easy to count and those that are difficult to count. This objective is met through over-sampling in harder to count areas (see also section 6 in this paper).

Stratification by LA was a new feature of the 2011 CCS design – in the previous CCS the main geographical stratification was done by estimation area, as the national sample size was shown to not support direct LA estimates in most cases (Abbott and Brown, 2008). **The 2011 approach to stratification worked well and the recommendation is to adopt a similar approach for 2021, that is, stratify the sample by LA and a 2021 version of the HtC Index.** The latter is being developed under a separate work package (Dini, 2018).

The HtC classification for the 2011 Census was developed using a model to predict area level non-response rates. The model predictions were used to rank all LSOAs in England and Wales from the lowest predicted non-response rates to the highest. This ranking was then used to split the distribution into five levels, from 1 (the least hard to count) to 5 (the hardest to count). The percentages of LSOAs in each category were: 40% in HtC1, 40% in HtC2, 10% in HtC3, 8% in HtC4 and 2% in HtC5. A similar approach was used for the design of the 2001 HtC index, which had three categories and used a 40/40/20% split.

The HtC index has traditionally been developed for use as a stratification variable in the CCS sampling and estimation, although in 2011 the index was also used to provide information for the census planning of field follow-up. The requirement for 2011 was to deliver a national HtC index with the following characteristics (Abbott, 2008a):

* A reasonably small number of levels (no more than 5) – depending on the confidence in the predictive power of the underlying data
* High correlation with the likely level of non-response in the 2011 Census
* Had to be at LSOA level or lower, depending on data availability
* Ideally using the most up to date data available
* Robustness - not knowing what the actual response patterns would have been in 2011, any decision taken in the development of the index had to be conservative.

In the 2021 Census, unlike previous censuses, the primary data collection mode will be an online (self-response) questionnaire. This means that in addition to the population groups that we know do not respond to a census, there will be groups of people who are digitally excluded and thus will require digital assistance or alternative ways of responding to the census that may not be via the primary mode. Thus, the 2021 Census requirements for the HtC index have been updated as follows (Dini, 2018):

* To plan where paper questionnaires are to be sent as first option
* To predict area response patterns at day 10 after the census (which will capture willingness to self-respond or not to a census, as field follow-up starts after day 10)
* To support where help or digital assistance will be required
* To support planning of follow-up resources, and their allocation
* To be a stratification variable for the CCS sampling
* To be used as required as a variable in the census estimation and adjustment.

As described by Dini (2018), the proposed 2021 HtC index is composed of two domains: a ‘digital’ domain and a ‘willingness to self-respond’ domain. For the digital domain LSOAs are ranked from those with the highest percentage of households having an internet connection to those with the lowest. For the willingness domain LSOAs are ranked from those with the lowest predicted non-response to a census by day 10 after census (as given by the model) to those with the highest. The ranked LSOA-level indicators in each domain are then split into 5 categories, from the least hard to count to the hardest to count. The percentages of LSOAs used in the split are: 40% in HtC1, 40% in HtC2, 10% in HtC3, 8% in HtC4 and 2% in HtC5. This gives rise to two separate HtC indicators of 5 categories each, and the two indicators can be combined according to different needs.

A decision has not been made yet on whether the digital domain component of the HtC index will be used for stratifying the 2021 CCS sample and in the census estimation. During the census operations all effort will be made to enable census responses from households that do not have access to the internet or the skills to respond to an online census. This will be done for example by sending a census paper questionnaire first instead of a letter with a Unique Access Code (UAC) to households that do not have access to the internet. In addition, during the census follow-up, support will be provided to households that do not have the skills to complete an online census. For this reason, it is important that the digital domain will be used to identify areas where we need to provide support to households that cannot complete an online census form.

However, we cannot rely on the full success of census strategies in this respect, thus, **we are considering stratifying the CCS sample by both the willingness and the digital domain**, or whatever other additional variables/data census will use to address the added risks of the new (for 2021) mode of collection. This is because we want to make sure that we will have enough sample in each stratum to be able to correct for possible biases induced by the effect of the new mode of collection in the census. **A final decision on the exact form of the HtC stratification of the CCS sample will be made after further research.**

Research was also undertaken to assess whether a model to predict final census non-response rather than day 10 non-response would be more appropriate to build a HtC index to be used for stratifying the 2021 CCS sample and in census estimation. The rationale for using a model to predict final census non-response would be that the CCS is conducted after the census field follow-up is completed. However, the results showed that there is no strong evidence to justify the use of two models, one to build a HtC index to aid planning of census field follow-up, and another to build a HtC index to be used in the CCS sampling and estimation (for more details, see Dini, 2018, section 3.3).

The index is required for both the 2019 Census (and CCS) Rehearsal and the 2021 Census (and CCS). The version used for the rehearsal may be used to validate and improve the index. The rehearsal version is required by November 2018. A date for when the 2021 version is required is not yet fixed.

# 6. Sample Size and Allocation

In 2011 the overall sample size for the CCS was approximately 17,400 postcodes covering nearly 340,000 households, although only 16,700 postcodes were used in the estimation process (ONS, 2012b). **For the 2021 CCS we recommend the same sample size as in 2011**, as the expected gains in precision from a larger sample are small compared to the likely increase in costs. In addition, anything larger than this would pose a significant risk to successfully delivering the CCS, which could then impact on quality (Abbott, 2008b).

**However, note that the overall number of households is only an approximate figure, as we will be sampling postcodes and not addresses** – the actual number of households to be visited will only be known after the postcode listing phase of the CCS (which will take place in the first two days of the fieldwork). Thus, we recommend that some contingency is built into the CCS fieldwork planning to consider this uncertainty, e.g., could work with ±10,000 households, rather than a fixed number.

Sample allocation refers to how the sample is allocated to the different strata (i.e., sample size per stratum) given a fixed overall size. Two typical sample allocation methods used in sample surveys are: proportional allocation and optimal allocation. Proportional allocation allocates the sample to strata according to the relevant proportions in the population (e.g., if stratum *s* makes up 2% of the overall population, then 2% of the sample will be selected from stratum *s*). Optimal allocation allocates the sample to strata according to a so-called ‘design variable’, which captures the population behavior with respect to whatever we are trying to measure (or with respect to a specific variable of interest) – in the case of the CCS this is mainly census undercount.

In 2011 optimal allocation was used in the first stage of sampling, albeit with the addition of some constraints (Abbott, 2008b, Abbott and Brown, 2008, Brown, Abbott and Smith, 2011, ONS, 2010). A design variable based on the 2001 household imputation counts (which was the best measure available of census undercount) was used to determine the PSU sample size within each stratum. The following constraints were also applied: a minimum of 1 PSU per stratum (i.e., per LA and HtC level) and a maximum of 60 PSUs in an LA. The first constraint ensured that the sample was well spread across geography. The second constraint prevented over-allocating the sample based on 2001 patterns. The design variable skewed the sample towards those strata where response was expected to be low or extremely variable, which tended to be urban and inner-city LAs where response was low in the 2001 Census.

The allocation used in 2011 was different from that used in 2001, which used less information, and different from that used during the CCS budget planning, which used a proportional allocation. The 2001 allocation was relatively uniform across all areas and groups, whereas the 2011 allocation put more sample into areas where undercount was expected to be high in 2011, based on 2001 Census data. Proportional allocation is somewhere in between the two, and moves the sample more towards the areas with bigger populations (typically urban areas), but not as much as optimal allocation, which skews the sample much more towards areas with lower expected census response (Abbott, 2008b).

In 2011 no attempt was made to use a design variable at the postcode level (i.e., in the second stage of sampling) as postcode level information tends to be much more unstable than information at higher levels.

The 2011 approach to sample allocation worked well. For a fixed cost, the optimal allocation approach offers gains in the precision of the estimates at a national level, while smoothing the precision across local areas, when compared to a proportional allocation and an allocation based on the 2001 methodology. Again, this was supported by simulation studies conducted in preparation for the 2011 Census (Abbott, 2008b). However, this is based on the assumption that the design variable is a good proxy measure for the variable of interest.

Thus**, the recommendation is to use a similar sample allocation approach for the 2021 CCS, including using a similar design variable with appropriate changes to guard against over-allocation**. So, for example, using the 2011 (rather than 2001) household imputation counts to determine the PSU sample size per stratum. **However, the 2021 approach should be more conservative,** that is, less reliance could be put this time on the 2011 patterns of non-response, as in 2021 the primary mode of collection of the census will be via online self-response, rather than via a self-response paper questionnaire. New simulation studies are currently being conducted to explore alternative sample allocation methods for the CCS using the improved estimation method that is under development for the 2021 Census.

# 7. Other Sample Design Questions

This section covers other CCS design-related aspects. These additional questions with provisional answers were also included in the 2021 CCS Requirements paper which defines what the CCS Field Operations team will deliver (ONS, 2018).

## 7.1. CCS Sampling Strategy for Communal Establishments

The CCS has traditionally only covered small communal establishments (CEs), which are defined as managed accommodation with between 10 and 99 bed spaces (ONS, 2018). Any managed accommodation with less than 10 bed spaces is treated as a ‘household’, and any with more than 99 bed spaces is treated as a large CE and is out of scope for the CCS. Large CEs are in scope for the census, and the coverage assessment and adjustment methodology for them does not use the CCS but alternative data sources (administrative data).

Two questions that have been raised at the 2021 CCS Working Group are:

1. for 2021 could we fix a sample size for small CEs in the CCS, so to get better estimates of the numbers of these CEs and their usual residents?
2. should we keep the same upper threshold of 99 bed spaces in the definition of small CEs in the CCS, or should we decrease this upper bound to 49 bed spaces, on the basis that in 2011 only a very small proportion of small CEs in the CCS happened to have between 50 and 99 usual residents?

The answer to the first question is that the CCS has traditionally been designed mainly to assess census coverage of households and persons within them. Small CEs are only covered by the CCS when they happen to fall in the sampled areas, and while estimates of the numbers of small CEs are reported separately from those of the numbers of households, estimates of the numbers of usual residents within small CEs are added to the estimates of the numbers of usual residents within households.

**Thus, fixing a sample size for small CEs to achieve a desired level of precision of the associated census estimates would require the development of a separate sampling strategy for them**. **This is possible; however, we think this is a question that should be addressed as part of the 2021 Census Coverage Assessment strategy,** including: what is the best strategy to assess coverage of, and within, communal establishments in the 2021 Census (including small and large CEs), and what is the impact of a change in the CCS to add a specific CE sample on the successful delivery of the survey (e.g., in terms of the field operations)?

More details about estimation and adjustment for CEs in the 2011 Census are available in ONS (2012d). The 2021 methodology of estimation and adjustment of CEs has not yet been fixed and is currently being reviewed.

**Regarding question b), or the definition of small CEs in the CCS, again this question should be addressed as part of the 2021 Census Coverage Assessment strategy, because if we change the definition in the CCS, this will need to be changed in the census as well.**

## 7.2. ‘Flexible Sample’ Option

The option of a flexible sample was used in the 2011 CCS. This involved the selection of some extra sample (across England and Wales), along with the main sample, to be used only in areas where during the census fieldwork the achieved response would be lower than expected. In 2011 the flexible sample was invoked on 15 April (three weeks before the CCS operational start date – 9 May) and resulted in the addition of approximately 7,000 addresses in 392 postcodes from 44 local authorities. The additional postcodes were supplied to Geography on 15 April 2011 for inclusion in existing CCS workloads (ONS, 2011, p. 68).

Although as mentioned in section 1 of this paper, boosting of the CCS sample in areas where census response was lower than expected did not go as well as was originally hoped (ONS, 2013, p. 15), this was mainly due to operational challenges. **Thus, we recommend that this option be considered again for the 2021 CCS and take on board the recommendation from the 2011 Census Evaluation Report** (ONS, 2013, p. 15), which recommendedthat boosting be considered much earlier in the planning process and consideration given to only use it if it can be shown to significantly improve the estimates. More work is required to understand what the recommendation from the evaluation report means in practice.

# 8. Summary of Recommendations for the 2021 Design

**Recommendation 1:** the CCS sample design is area based, so postcodes rather than addresses are the final sampling unit.

**Recommendation 2**: Keep the 2-stage design approach, and keep output areas as the PSUs, while consideration should be given to reducing the SSU sampling fraction from the 2011 level of 50% to 25%. The analysis of the precision of the census estimates and of the costs support the use of this reduced sampling fraction; however, a smaller scale simulation study is recommended to ascertain that the results still hold when using the improved estimation method under development for the 2021 Census.

**Recommendation 3**: Keep the same stratification variables as in 2011, that is, LA and an updated version of the Hard to Count Index. The updated version of the HtC index to be used in the CCS sampling is the same index that is being developed for the planning of the 2021 Census field follow-up. However, afinal decision on whether we will be using both the ‘willingness’ and the ‘digital’ domain of the 2021 HtC index, or only the ‘willingness’ domain, has not been made yet and requires further research.

**Recommendation 4**: Keep the overall sample size (i.e., overall number of postcodes and households) the same as in 2011; however, for CCS field-work planning purposes the suggestion is to work with a range of (for example) ±10,000 households, rather than a fixed number.

**Recommendation 5**: The sample allocation method is like the 2011 approach; however, further analysis is recommended to assess the option of using a method that is less reliant on information about the 2011 Census coverage patterns.

**Recommendation 6**: The question on the definition of small CEs and the question of whether it is possible to set a target sample size for them in the CCS should be addressed as part of the 2021 Census Coverage Assessment strategy, as the answer to these questions is driven by Census requirements for measuring coverage of CEs (both large and small).

**Recommendation 7**: For the 2021 CCS consider the option of a flexible sample again, and take on board the recommendation from the 2011 Census Evaluation Report, that is,that boosting be considered much earlier in the planning process and consideration given to only use it if it can be shown to significantly improve the estimates. More work is required to understand what the recommendation from the evaluation report means in practice.

# References

Abbott O. (2008a) Requirements for a Hard to Count index. Internal ONS document.

Abbott O. (2008b) CCS Sample Size and Allocation. Internal ONS document.

Abbott O., and J. Brown (2008) CCS Design Strategy for the 2011 Census. Internal ONS document.

Brown J., O. Abbott and P. A. Smith (2011) Design of the 2001 and 2011 Census Coverage Surveys for England and Wales, *Journal of the Royal Statistical Society A*, 174(4), pp. 881-906

Castaldo A., and O. Abbott (2016) Design of the 2021 Census Coverage Survey. Paper prepared for the Census Transformation Programme Research Assurance Group, March 2016

Castaldo A. (2017) The 2021 Census Coverage Survey: the option of moving to an address based sampling design – Preliminary evidence. Paper prepared for the Census Transformation Programme Research Assurance Group, January 2017

Castaldo A., and A. Harfoot (2017) The 2021 Census Coverage Survey: using AZTool to estimate costs for interviewers. Presentation prepared for the 2017 RSS Annual Conference, Glasgow, 4-7 September 2017. Available at <https://events.rss.org.uk/rss/media/uploaded/EVRSS/event_93/1.2_An_innovative_approach_to_estimating_the_costs_of_the_2021_Census_Coverage_Survey.pdf>

Castaldo A., and D. Nikolakis (2018) Assessing the Use of an Address Based Design for the 2021 Census Coverage Survey. Version 4, 17 May 2018.

Castaldo A., S. Cockings and A. Harfoot (2018) Using the AZTool to estimate the costs of the Census Coverage Survey. Version 4, 13 September 2018.

Cockings S., A. Harfoot and D. Martin (2017) Automated zone design for data collection: designing workload areas for censuses and surveys. Presentation prepared for the Census Research User Conference, RSS, London, 22 June 2017. Available at <https://www.ukdataservice.ac.uk/media/604839/cockings.pdf>

Dini E. (2018) Hard to Count Index for the 2021 Census. ONS, Methodology, April 2018.

Hill M., and O. Abbott (2010) 2011 Census Coverage Survey – creation of the sample and frame. Internal ONS document.

Hopper N. (2011) Predicting patterns of household non-response in the 2011 Census. *Survey Methodology Bulletin*, 69, pp 9-22. Available at <http://www.ons.gov.uk/ons/guide-method/method-quality/survey-methodology-bulletin/smb-69/survey-methodology-bulletin-69---jan-2012.pdf>

ONS (2010) 2011 Census Coverage assessment: CCS sample sizes and Estimation Areas for local authorities – autumn 2010. Available at

<http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/census-consultations/uag/census-advisory-groups/statistical-development/census-coverage-assessment---ccs-sample-sizes.pdf>

ONS (2011) CCS Operations Evaluation Report, FOPS.CCS.0089.v1.0.

ONS (2012a) 2011 Census Coverage Survey - 2011 Census Evaluation Report. July 2012. Available at

<http://www.ons.gov.uk/ons/guide-method/census/2011/how-our-census-works/how-did-we-do-in-2011-/evaluation---census-coverage-survey.pdf>

ONS (2012b) 2011 Census Coverage Survey Summary - 2011 Census: Methods and Quality Report. October 2012. Available at

<http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/census-coverage-survey-summary.pdf>

ONS (2012c) Census Coverage Survey Sample balance adjustment. July 2012. Available at <http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/ccs-sample-balance-adjustment.pdf>

ONS (2012d). Estimation and Adjustment for Communal Establishments - 2011 Census: Methods and Quality Report. September 2012. Available at: [http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/coverage-within-communal-establishments.pdf](http://webarchive.nationalarchives.gov.uk/20160112053850/http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/coverage-within-communal-establishments.pdf)

ONS (2013) Coverage Assessment and Adjustment Evaluation - 2011 Census Evaluation Report. March 2013. Available at

<http://www.ons.gov.uk/ons/guide-method/census/2011/how-our-census-works/how-did-we-do-in-2011-/evaluation---coverage-assessment-and-adjustment.pdf>

ONS (2018) 2021 Census Coverage Survey Requirements. Paper submitted by Methodology and Census Statistical Design for the CCS Field Operations. Final version, June 2018, v5.1.

Ordnance Survey (2014) AddressBase: the product name for the National Address Gazetteer in which address information from the Royal Mail, Local Authorities and the Ordnance Survey have been brought into one database by the GeoPlace organisation. Available at <http://www.ordnancesurvey.co.uk/business-and-government/products/addressbase-products.html>

1. Local Authorities are grouped together into estimation areas, so that there is sufficient sample to reliably estimate the population in those areas. For more information see ONS (2010). [↑](#footnote-ref-1)