United Nations ECE/CES/GE.41/2015/17

Economic and Social Council

Distr.: General
3 September 2015

English only

## Economic Commission for Europe

Conference of European Statisticians

**Group of Experts on Population and Housing Censuses**

**Seventeenth Meeting**
Geneva, 30 September to 2 October 2015
Item 3 of the provisional agenda
**Innovations planned for 2020 census round, and results of tests**

# Research for 2021 Census England and Wales: possible innovations under consideration

## Note by the Office for National Statistics, United Kingdom

*Summary*

For the 2021 Census the Office for National Statistics(ONS) plan to deliver a predominantly digital census while making the most effective use of administrative data in its design, operation and outputs. The main areas of change that ONS are researching for the 2021 design include: designing for and accessing an online questionnaire first (rather than paper); devising new strategies for identifying areas and populations that are hard-to-count and/or digitally excluded; increased use of administrative data within various aspects of the design, operation and statistical processing. This paper will summarise early research into some of these aspects. The electronic version of the paper includes a number of hyperlinks

# I. Introduction

1. This paper summarises early research behind some aspects of the design of the 2021 Census in England and Wales.

2. The paper highlights the drivers for change for the 2021 Census Transformation Programme that have helped shape the initial design, which draws on experiences from the 2011 Census, current world best practice, and consideration of technological and societal changes. The 2021 Census Design Document sets out a fuller picture on our initial thinking of the entire design of the 2021 Census.

3. The ideas presented will be subject to further research, building on feedback received during consultation and advisory meetings and a detailed programme of testing over the coming years. In addition, the final design will also be informed by work to understand the cost-benefit trade-offs followed by commitment to resourcing the programme for 2021.

# II. Drivers for change and design principles for 2021 Census design

4. Having extensively reviewed all the available options and evidence through the Beyond 2011 programme – including international comparisons, statistical research, public attitude research, responses to public consultation, cost-benefit analysis and independent review – in March 2014, the National Statistician recommended:

- An online census of all households and communal establishments in England and Wales in 2021 with special care taken to support those who are unable to complete the census online
- Increased use of administrative data and surveys in order to enhance the statistics from the 2021 Census and improve annual statistics between censuses.

5. This will make the best use of all available data to provide the population statistics which England and Wales require and offer a springboard to the greater use of administrative data and annual surveys in the future.

6. The initial high-level design for a 2021 Census is informed by:

- Lessons learned and successes, where appropriate, from the 2011 Census
- Developments and lessons learned in international census taking
- Requirements from the user community about the types, quality, frequency and detail of outputs required
- Changes in technology, in particular the opportunities offered by the internet, and changes in the propensity for the public to interact with government, enabling a move away from a traditional paper based census
- Improvements in administrative data sources (such as the patient register and tax and benefits data) giving potential for increased use in the production of statistics
- The continued and on-going need to make the most effective use of public money

7. The design will evolve as the programme develops and the different design elements are first tested independently, and then tested together. This paper summarises some of the aspects being tested, which are a small part of the whole testing strategy being developed,

which includes ongoing methodology research, small scale and large scale testing and rehearsals.

# III. Ongoing research on selected key elements of the design

8. The move to a predominantly online census has many opportunities for all aspects of the 2021 Census, from the design of the questionnaire, to the management of the field operation, to the processing and production of outputs. This section provides an overview of some of the research that is ongoing in these areas.

## A. The online questionnaire

9. The 2021 Census will be primarily an online census. Building on the successful 2011 online questionnaire design, research is being undertaken to understand how to transfer the questions to mobile and electronic devices.

10. The online questionnaire will take account of best practice standards and guidance, and will be monitored throughout the development phase to ensure compatibility with the most common browsers, devices and operating systems.

11. In addition, the online questionnaire will be designed independently from any paper questionnaire to maximise online take up and data quality. Opportunities that will be considered and tested include:

- Contextual help to help complete questions
- Use of detailed drop down boxes to reduce, or eliminate altogether, the amount of coding for more detailed questions like occupation, industry or country of birth
- More comprehensive validation within and between questions
- Design of questions to fit smaller screens

12. Work will be needed to assess possible mode effects due to different designs between paper and online.

## B. Target population and digital inclusion

13. A primarily online census introduces new challenges. There are always likely to be hard-to-count populations who are at risk of low levels of engagement or response. These are by their nature specific and require special attention. These groups were identified and prioritised in the 2011 Census, and this work will be built on for 2021, to understand their needs and explore appropriate ways to achieve higher levels of response. Digital exclusion is an additional factor for 2021, and typically affects some of the most vulnerable and disadvantaged groups in society. For instance, an estimated 10 per cent of the adult population (not households) may never be able to gain basic digital capabilities because of disabilities or basic literacy skills (see Cabinet Office Report). The design and operation of the census will therefore need to take particular account of the requirements of these individuals.

14. To encourage online response, we need to ensure that services are in place that take account of respondents who would like to complete online but are unable to. Understanding our respondents and how they wish to interact with a census collection exercise, based on an understanding of interactions with other government services, is key to achieving the required response rate. Current research takes account of assisted digital

requirements to meet Government Digital Service (GDS) guidelines. Information from recent surveys and censuses has identified four groups of respondents to an online census, split between those who are willing to complete online and those able to do so, as shown in Figure 1.

Figure 1
**Matrix with the four groups of respondents to consider in an online census**

| | No access/Do not use internet | Access and use of internet |
|---|---|---|
| Willing to use the internet to complete government processes online | **Group 2** | **Group 1** |
| Not willing to use the internet to complete government processes online | **Group 4** | **Group 3** |

15. Evidence is being gathered from surveys and censuses to try and quantify and identify characteristics of people within these four groups. For instance:

- the assumption is being made that information from surveys on peoples' use of the internet for digital government services is a reasonable proxy for responses in Group 1. More UK government services are moving to digital by default design, so more information will become available on the demographic characteristics of this group in the future.
- people in Group 2 are more likely to be in older age groups, lacking access to the internet and/or the skills to use it. There may also be other barriers such as financial issues.

16. However, research needs to be done to assess peoples' willingness to respond online. Groups 3 and 4 will be an extension to the simply 'unwilling' to contribute groups in previous censuses (some characteristics of which will be available from 2011 response rates), as the added concern around internet security may influence views on responding online. Finding the characteristics of people across these two groups will be more problematic, but could to some extent be predicted by finding out about those not engaging with other government digital services due to trust issues, if this was possible.

17. Research on the characteristics of internet respondents is ongoing, controlling for key characteristics of responders by each mode in order to detect any underlying differences, using a Propensity Score Method (Fraser & Ghee, 2015), a method also used by Statistics Canada on the 2006 Canadian Census. We have only been able to assess the characteristics of those who chose to respond online as opposed to what the situation will

be in 2021 where the online option will be pushed as default, but this research has demonstrated that some apparent differences in respondents by mode are not necessarily significant once other key characteristics are taken into consideration. For example, while initial summaries indicate that people with disabilities are less likely to respond online, this analysis implies that this effect could largely be driven by the propensity of people with disabilities to also be in the older age categories or have other factors that related to choosing to respond on paper. Thus the difference in response channel by disability when adjusted was actually close to zero. Further work is needed to refine the modeling behind the method, leading to better understanding of what determines preferred response mode, therefore contributing to our understanding of the groups who may need more support in 2021.

## C. Use of a hard-to-count index in design and prioritisation

18. In the 2011 Census, the strategic aim was to maximise overall response but also minimise the variability in non-response (Abbott & Compton, 2014). This resulted in the initial collector resource allocation being targeted at areas where the initial return was predicted to be lowest. This allowed the movement of those resources to focus on areas where the actual return rates during the field operation were lowest. A hard-to-count index was used to drive this allocation and prioritisation. The index was based on external data for areas containing around 750 households. For 2021, the strategic aim is the same. However, there are likely to be data that could support lower level targeting, potentially at address level. This could enable a more efficient targeting of resources, provided that predictions of likely non-response patterns are accurate. Research is underway to identify such data and assess its ability to predict non-response patterns. However, as noted above, predicting the patterns of non-response may be more of a challenge than in 2011 as a result of the online first approach.

## D. Field force management

19. The objective of the field design will be to maximise overall response, whilst achieving sufficient response in all local authorities to enable successful estimation of the population. To achieve this we need to track responses to understand response patterns and, using area based response targets identified by the hard-to-count work, allow us to direct follow-up activities. Research will be done to test early analysis of online responses - comparison of distributions of key characteristics of responding households and people with expected distributions, enabling flexible deployment of the field force in areas where distributions are not as expected. This will be explored in conjunction with work to design the field operations to enable this flexibility, recognising that there will be some constraints to how far field resources can be moved or flexed. The ongoing work on distinguishing the expected characteristics of responding households in the Hard-to-count and Propensity Score methods mentioned above will feed into this.

## E. Follow-up of non-responding householders

20. As with previous censuses, we will use different strategies to contact non-responding households and encourage them to take part. The precise timing and combination of follow-up modes is under development, but it will include the use of reminder letters and household visits from field staff. Use of telephone and email follow-up is also under consideration. As in 2011, there will be support available primarily online, and also in local venues such as libraries, drop-in centers and religious centers.

21. A field operation simulation model is being developed, using operational research techniques, to test various follow-up scenarios. This tool can be used to help decide the optimal follow-up approach, balancing quality objectives against the costs associated with the effort to achieve these objectives. The model uses intelligence from the previous census and can be tuned to area-specific inputs. Research into the model is at an early stage, but a simple model has been developed that uses data on:

- Number of households N
- First reminder letter posted at day X of the collection period
- Second reminder letter posted at day Y
- Impact of reminder letters
- Follow-up starting at day Z
- Maximum number of visits per household M
- Impact of a field visit
- Digital exclusion level D
- Response probability of digitally excluded households R

22. The output of the model includes

- Final return rate
- Returns by internet
- Returns by paper
- Whether contact was made, number of follow-up visits necessary
- Estimated costs

23. Figures 2 and 3 demonstrate the response profiles of two areas with all parameters similar, except the first has a digital exclusion level of 10%, the second 50%. The final response rate was similar in both areas (97.4% and 98.6%), but the figures show clear differences in the number of households responding by internet and paper. (It should be noted that the model has random factors at work, and results should be summarised over a number of simulated runs).

Figure 2
**Low digital exclusion**

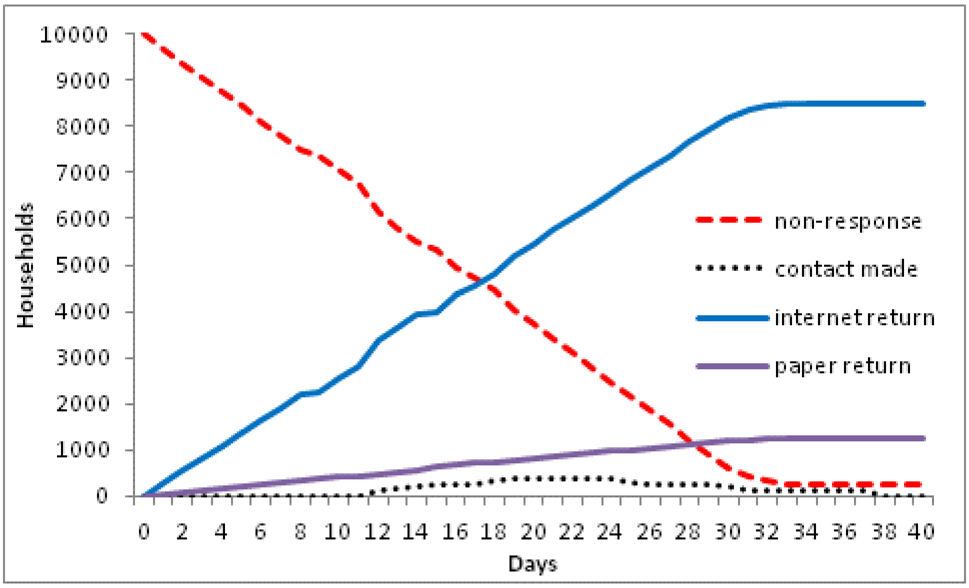
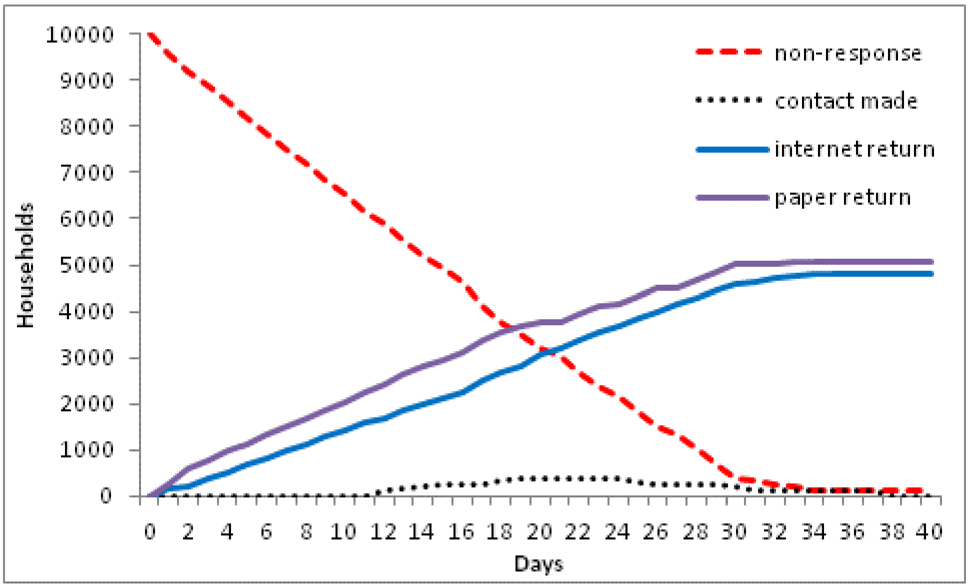


Figure 3
**High digital exclusion**



## F. Alternative census coverage methods

24. The 2011 Census used a Dual System Estimation (DSE) approach (Brown, 1999) alongside a ratio estimator to estimate the total population. However, DSE can be problematic as the assumptions underpinning the method are often violated. The requirement for very high precision matching can be particularly difficult, and costly. In 2011 a number of methods were implemented to measure and make adjustments. While successful, these were challenging to develop. For 2021, a weighting-class approach is being considered as an alternative to DSE (as described by Abbott et al (2015)). For this

approach, high quality individual level linkage is not required as linkage is done at the address level. The estimator is also less susceptible to over-coverage in the census. More research is being undertaken before any decision is made as to which estimator to use in 2021.

25. With the changes to the way the data will be collected in 2021 (mainly online, so with minimal if any scanning required), potentially removing the need to process on an area-by-area basis, there is the possibility of developing an estimation model for the whole country. This will enable the reduction of bias by refining DSE stratification and increase precision by increasing effective sample sizes. This will require further research.

## G. Use of administrative data in the census design – improving census coverage

26. A significant change from 2011 will be the greater use of administrative data in the design and conduct of the 2021 Census. Administrative data could be used to improve the quality of the census estimates. In the 2011 Northern Ireland Census, high quality administrative data was added to the census data for some households that did not respond. This improved the quality of the resulting statistics. Such an approach would depend on the quality and availability of suitable administrative data and the outcome of further research to understand the implications and quality gains of such an approach.

27. The availability of a record level administrative dataset for the whole country can be used to create a combined list of the population by matching it to the census to identify records on the administrative data that were not included in the census. These records can be used to augment the census. This augmented list can then be matched, in the sample areas, to the census coverage survey allowing dual system estimation to be performed. The assumption is that different types of people are missed by either the census or the administrative list, so combining them yields greater population coverage. This was the method used by Northern Ireland in their 2011 Census Under Enumeration (CUE) project, which resulted in an additional four per cent of individual records added to the census database (NISRA, 2014).

28. The method relies on the assumptions of a high quality address register, perfect matching and a robust methodology. The use of activity based administrative data in census enumeration was unique to Northern Ireland and research carried out by Ross (2015) showed that the project was highly successful, significantly improving the overall quality of the population estimates through decreasing their variance (see Table 1 below).

Table 1
**Estimates, Variances and Confidence Interval Widths by Estimation Area**

| *Estimation Area* | *With CUE records* | | | *Without CUE records* | | |
|---|---|---|---|---|---|---|
| | *Estimate* | *Variance* | *Relative CI width* | *Estimate* | *Variance* | *Relative CI width* |
| Eastern Northern Ireland | 791,900 | 9,146,600 | 0.75% | 782,000 | 14,372,000 | 0.95% |
| Western Northern Ireland | 535,800 | 10,800,700 | 1.20% | 528,600 | 13,991,800 | 1.39% |
| Belfast | 462,000 | 11,562,800 | 1.44% | 452,500 | 21,540,600 | 2.01% |

29. This research has demonstrated how a similar approach could be implemented in England and Wales using simulation studies to analyse the effect of the addition of administrative data on the quality of the estimates, and the dependence of the results on the census response rate. Due to the absence of an activity variable on the administrative data sources available at the time of the research, the studies used the patient register matched to the 2011 Census to model the likelihood of an individual being present at the correct address. The results showed that the use of administrative data as part of the coverage adjustment process in England and Wales improved the quality of census population estimates. One of the most important findings from the research was the significant improvement in the quality of the estimates when administrative data were used alongside census data which had a similar response rate to that achieved in 2011. This suggested that had this method been employed in 2011, the quality of the census population estimates could have been improved.

30. Despite the success of this method, using administrative data to compensate for non-response in the census requires the following:

- A high quality administrative data source with an indicator of activity.
- A cautious approach in adding administrative records to prevent overcoverage.
- High quality matching between the census addresses and the administrative records.

31. The research showed that the use of administrative data in this way had the potential to improve the quality of the census population estimates and recommended the method should be considered in the design of the 2021 Census (Ross, 2015).

## H. Improving post-coverage whole-record imputation

32. The basic record-level imputation methodology used in the coverage adjustment in the 2011 Census worked well to provide a database that was fully adjusted to take account of the measured coverage, adding wholly missed households and persons within existing households. It was based on the methodology described by Steele et al (2002). However, the implementation of the methodology was challenging. The main issue was with the calibration process which derived the household weights for imputing wholly missed households (and the people within them). This step did not always work and required collapsing or removal of variables in order to converge. Another issue with the adjustment system was that the donor imputation could use a record multiple times, causing spikes in benchmarked or non-benchmarked variables.

33. For 2021 further research into the coverage adjustment methodology will be carried out assessing both improvements to the 2011 methodology as well as looking at a range of alternative options. One of these options is a combinatorial optimisation approach. This involves finding the optimal combination from a finite set of possible solutions to a problem, within the constraints given by the coverage estimation stage. The approach essentially re-weights households with integers which are initially set to zero, and increased when a combination satisfies the required constraints.

34. Research so far has highlighted issues to be resolved, such as the presence of high weights (some households end up being used a large number of times), and has only been tested on one estimation area. However the method appears to have the potential to address the main issues found with the method used in 2011.

## I. Synthetic data for testing and outputs

35. A common constraint brought up in 2011 evaluations was the lack of adequate data for testing that sufficiently mirrored the complexity of census data. In order to address this, a number of synthetic data products are being developed across various parts of the programme, including data for matching and microdata products.

### 1. Matching

36. Synthetic datasets have been developed including variables required for matching, by an algorithm that is able to simulate census, survey or administrative data. The method makes use of 2011 Census patterns of names – first names by age, surnames by geographic region – in order to better replicate real-life situations. In order to reduce disclosure risk, census distributions have been used, but the real names replaced randomly from other name lists. The algorithm is able to create large sets of test data, also including random features that could occur in real data – for example naming differences, errors in date of birth entry or address spelling errors.

### 2. Microdata

37. ONS have provided some microdata products for 2011, but users are still asking for more. For 2011 Census we have provided a household microdata sample within our secure Virtual Microdata Laboratory (VML), but no other household set. We have provided individual microdata samples in the VML, via a licensed route (the UK Data Service) and a public teaching file on the ONS website. The Market Research industry and demographers have expressed a need for 'a more accessible version' of household microdata, with adequate variable and geographic detail.

38. The issue around the provision of household microdata is there is a high risk of disclosing personal information. For instance, for households with more than five people, the combinations of age and sex in the households can be unique at a relatively high geographic level. Adding in other census characteristics increases the identification risk for all households.

39. The method being tested punches holes in a microdata sample, and uses the edit and imputation process from 2011 to fill in the holes, thereby attempting to preserve the relationships between variables within households and individuals, but introducing sufficient uncertainty to enable greater access to the sample by users. We are testing different hole-punching scenarios to ensure we not creating sample bias. Risk is evaluated by examining unique records and will possibly do some intruder testing. Utility will be evaluated by comparing distributions, marginal totals and key cross-tabulations before and after perturbation.

40. While over-imputation was rejected as a method of disclosure control in 2011 Census, what is being trailed here is making use of the parameterization set up in CANCEIS to replicate underlying patterns in the data during item level edit and imputation, as long as there is sufficient information left in each record to find a suitable donor that fits the underlying patterns we specify.

41. We are looking at what amount of hole-punching is necessary in a 20% sample to maintain enough utility while leaving negligible risk (for a public file) or a small amount of risk (for a limited access household file to complement the individual file already available to registered users).

# IV. Summary

42. The above summarises a few of the innovations currently being developed and tested for the 2021 Census. The Census Transformation Programme is developing a testing plan, including successive small scale tests and a large scale test 2017 covering all aspects of the census process. In addition, research is ongoing on statistical processes building on 2011 experiences, and new approaches can be tested using the synthetic data sets being developed. Results of the testing and updates to the census design will be published as they become available.

# References

Abbott, O., Castaldo, A., Racinskij, V., Ross, H., Smith, P. and Brown, J. J. (2015) Developing a weighting-class approach for the 2021 Census. Government Statistical Service Methodology Advisory Committee Paper 29/3. Available at http://www.ons.gov.uk/ons/guide-method/method-quality/advisory-committee/previous-meeting-papers-and-minutes/mac-29-papers.pdf

Abbott, O. and Compton, G. (2014) Counting and estimating hard-to-survey populations in the 2011 Census. In: R. Tourangeau, B. Edwards, T. Johnson, K. Wolter and N. Bates, eds. 2014. Hard-to-survey populations. Cambridge: Cambridge University Press. Ch.4.

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.

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. Journal of the Royal Statistical Society A, 162, 247–267.

Cabinet Office, (2014), Government Digital Inclusion Strategy,

https://www.gov.uk/government/publications/government-digital-inclusion-strategy/government-digital-inclusion-strategy

Fraser, O. and Ghee, C., 2015. Analysis of the characteristics of internet respondents to the 2011 Census to inform 2021 Census questionnaire design. Twentieth GSS Methodology Symposium, London. 1st July 2015.

Grondin, C. and Sun, L., (2008). 2006 Census Internet Mode Effect Study. American Statistical Association Joint Statistical Meeting, Section on Survey Research Methods. 3-7 August, Denver. Available at https://www.amstat.org/sections/SRMS/Proceedings/y2008/Files/300977.pdf

Nowok, B, Raab, G. M. and Dibben, C. (2015) Synthpop: Bespoke Creation of Synthetic Data in R. CRAN vignette. Available at www.cran.r-project.org/web/packages/synthpop/vignettes/synthpop.pdf

Office for National Statistics, (2015), 2021 Census Design Document. Available on request.

Ross, H., 2015. Using administrative data to enhance the quality of census population estimates. MSc.Thesis, University of Southampton.

Steele F, Brown J and Chambers R (2002) A controlled donor imputation system for a one-number census. Journal of the Royal Statistical Society A. 165, 495–522.

UK Statistics Authority, 2014, The census and future provision of population statistics in England and Wales, Available at http://www.statisticsauthority.gov.uk/news/statement---census-and-the-future-provision-of-population-statistics-in-england-and-wales---27032014.pdf