2011 Census Data Quality Assurance Strategy

CONTENTS

Executive Summary

1 Introduction

1.1 What is quality and how is it managed in the 2011 Census?
1.2 The objectives of the 2011 Data Quality Assurance Strategy
1.3 Guiding principles for the 2011 Data Quality Assurance Strategy

1.4 Scope of the Data Quality Assurance Strategy
1.5 Components of the Data Quality Assurance Strategy and their timing
1.6 Interfaces and dependencies with other data systems
1.7 Improvements in data quality assurance since 2001
1.8 Skills and roles required for data quality assurance

2 Achieving Our Quality Objectives

2.1 Quality assurance of census topics (Topic QA).
- 2.1.1 The variable checking framework
- 2.1.2 Comparator data and statistical tolerances
- 2.1.3 Change over time in the topic QA strategy
- 2.1.4 Topic QA at regional and national levels
- 2.1.5 The role of topic experts
- 2.1.6 Interface with population counts and structures: population subgroups
- 2.1.7 Modal effects: Monitoring the difference in paper and internet data

2.2 Quality assurance of population counts and structures (Demographic QA)
- 2.2.1 Quality assuring the estimated population totals through quantitative and qualitative analysis
- 2.2.2 Quality assuring the post coverage imputation census database
  - 2.2.2.1 Population subgroup analysis
  - 2.2.2.2 Cohort analysis
  - 2.2.2.3 The address register as a QA source
- 2.2.3 Reconciliation of national and local estimates
- 2.2.4 Evidence of multiple enumeration and adjustments for usual residence
- 2.2.5 Local authority strata and the checking framework
- 2.2.6 Key data sources and statistical tolerances
  - 2.2.6.1 Assessing quality of comparator data
  - 2.2.6.2 Methodological enquiries to inform the use of comparator data
  - 2.2.6.3 Methodological enquiries around particular population subgroups

2.2.6.4 Statistical tolerance for LADs
2.2.7 Supporting evidence
2.2.8 The relationship between demographic QA and the census coverage adjustment
2.2.9 Quality assurance and the 2011 Census quality survey
2.2.10 Stakeholder input
2.2.11 Outcomes of the QA Panel
2.2.12 Where agreement of estimates cannot be achieved: supplementary analysis

3 Transparency, Peer Review and User Communications

3.1 Quality Assurance Panel
3.2 Peer review
3.2.1 Quality Assurance Working Group (QAWG)
3.2.2 UK Census Design Methodology Advisory Committee (UKCDMAC)
3.2.3 International peer review
3.2.4 Harmonisation across the UK
3.3 User communication strategy
3.4 Outputs from the data quality assurance process

4 Data QA Strategy for the 2009 Rehearsal

5 Next Steps

6 Bibliography

Appendix A Lessons learned from 2001
Appendix B Data quality monitoring system: interfaces with other data producers and systems
Appendix C Prioritised data quality assurance tasks
Appendix D Indicative checking framework for topic QA
Appendix E Sources of data for population subgroup analysis used in 2001
Appendix F Multiple enumerations in 1991
Appendix G Datasets used for 2001 quality assurance
Appendix H Additional datasets to be considered for 2011 quality assurance
Appendix I Data QA checks in the 2009 Census Rehearsal
Appendix J Timeline for data quality assurance activities

Executive Summary

The 2011 Data Quality Assurance (QA) Strategy has been produced following a review of literature including the 2001 Census Quality Strategy, comments following the 2001 Census and examples of international best practice. Interviews with people involved in the 2001 Census have also ensured that the 2011 strategy reflects lessons learned from the last census. It has been guided by information from the UK Quality Assurance Working Group and the UK Census Design Methodology Advisory Committee, both involving subject experts and census managers in Northern Ireland and Scotland and representatives from the Welsh Assembly Government. To check robustness, the QA approach proposed for 2011 has successfully identified the local authorities that posed challenges in 2001 and where census estimates were queried.

The Data Quality Assurance Strategy should be understood in the context of the 2011 Quality Strategy, which describes how the aims of the 2011 Census will be monitored, managed and met. The strategic aims are:

- to provide high quality statistics that meet user needs
- build confidence in the final results
- provide value for money solutions
- to protect, and be seen to protect, confidential personal census information

Critical success factors will give tangible evidence of whether those aims have been met. Quality throughout the census will be managed using a quality model. It has five key strands:

- design quality
- operational quality management
- data quality assurance
- quality measurement and reporting
- project quality management.

The Data Quality Assurance Strategy contributes primarily to the third and fourth strands.

Quality assurance will start in the two to three weeks before census day, when management information and early census returns will provide early evidence of response patterns and characteristics, and will continue through to the publication of outputs. The QA activities described in this strategy represent what could be done to validate census results. A process of planning and prioritisation will determine where QA resource is focused. Early research findings could also shift the focus of implementation.

Two distinct but complementary strands of data validation will be conducted in parallel, referred to as ‘topic’ and ‘demographic’ QA. Topic QA involves validating census data and counts as they pass through the processing

systems, using comparator datasets and monitoring changes to data distributions. Demographic QA involves validating, adjusting and accepting census population estimates at local authority district level, following adjustment for under- and over- coverage. Both topic and demographic QA will get support from subject experts and both will assess distributions of population subgroups posing known enumeration challenges such as immigrants, armed forces, students, babies under a year old and young men. Topic QA will seek the early identification of data anomalies so that adjustment can be made to the relevant systems or processes. Demographic QA will necessarily be focused on census estimates following coverage adjustment and will involve intense input from the quality team and a QA Panel. The QA Panel will have responsibility for recommending local authority, regional and national census population estimates for ONS executive sign-off.

Topic QA will involve examining data before and after key processes including data capture (for internet responses), the reconciliation of within-household multiple responses, item imputation, coverage imputation, derived variables and allocation of output geographies. As well as comparing pre- and post-process distributions against expected values, cross-tabulations will ensure the internal integrity of the data. Experts in the fields of demography, employment, education, health, housing and identity will guide and inform data validation. Checks will be carried out on variable distributions at the levels of local authority, region, national and cumulative total.

The quality team will build expected census population estimates using a range of administrative data sources and rolled-forward ONS mid-year population estimates. The census estimates for each local authority district, region and country will be compared against these and adjustments made as appropriate. A series of demographic indicators will provide further validation, for example sex and dependency ratios. Local authority counts and cumulative totals and cumulative distributions of key population subgroups will be monitored against non-census sources. A QA Panel of expert demographers and representatives from the Local Government Association and the Welsh Assembly Government will make recommendations for each local authority district estimate. The QA Panel will operate two parallel subgroups, one ‘priority’ group focusing on the local authorities posing particular challenges for enumeration or coverage adjustment, and the other subgroup approving all remaining estimates. A working assumption is that around 25 per cent of local authority districts will be reviewed by the priority subgroup.

Where there is an unexplained discrepancy between expected values and the census estimates, the QA Panel will draw on further evidence provided by supplementary analysis which is likely to include use of administrative records. For the 2001 Census ONS carried out a series of studies to improve population estimates in the areas that were hardest to count. This included address matching studies in two authorities, Manchester and Westminster, using administrative address lists from their city councils and the address list collated by the ONS for the 2001 Census. In 2011, the strategy involves

bringing this process forward so that new evidence on problem subgroups or areas will begin in 2010 and be available to inform the decisions of the QA Panel in 2011/12. Where possible, aggregate-level data will be used. Micro-data matching from different sources will also be considered, pending legal gateways including parliamentary data sharing orders if necessary.

Trust and confidence in census data will be supported by a transparent approach to the methods and results of the QA processes. The Quality Assurance Stakeholder Communications Strategy identifies a wide range of stakeholders, including local authority representatives and other government departments (local and national), academia, professional bodies and international partners. It is envisaged that input from local authority representatives will be coordinated at a regional level through ONS regional statisticians. Comments and suggestions for alternative comparator sources will be actively sought and all QA decisions will be published on the internet. Key deliverables from the QA strategy will be high quality outputs accompanied by timely metadata in the form of data quality reports and a report for each local authority district.

Prioritisation will ensure that QA activities are delivered on time. Topic QA will focus on validation of early data batches, delivered prior to processing in the case of internet data capture. It will prioritise variables used to generate census population estimates, directly or indirectly - for example those informing coverage imputation. Management information from the field will signal likely data quality issues and information gaps. Demographic QA will prioritise known problematic subgroups and local authorities. Supporting evidence from administrative sources, including data matching where necessary and involving evidence from the ONS longitudinal study (LS), will identify in advance and fill known information gaps arising from field operations or as a result of QA checks. QA activity will be pre-programmed as far as possible so that analytic expertise can focus on data anomalies identified through automatic screening checks. This work will be supported by data visualisation and, possibly, spatial analysis techniques. The ONS executive, considering the recommendations of the QA Panel, will have the options of: i) approving estimates, ii) commissioning further investigations/ analysis, iii) requesting data matching for further evidence or iv) rejecting estimates, even if this means that publication of results will be delayed.

1 Introduction

This strategy describes planning currently underway to quality assure (QA) 2011 Census data. It provides a general framework for developing and carrying out processes to define, assess and manage the quality of 2011 Census data. Lessons learned about quality from the 2001 Census have informed this strategy and are described in Appendix A.

The strategy addresses assurance of census counts and census population estimates. A census coverage survey will be conducted six weeks after census day. This involves an independent count of all households and individuals in a sample of postcodes. The survey will be designed to estimate the level of undercount in the census, allowing an assessment of the characteristics of individuals and households missed. Census counts will be adjusted to take account of the missed population, with generalisation to the areas not covered by the coverage survey. Households and people estimated to have been missed will be imputed into the census database. Census estimates, including the adjustment for under-enumeration, need to be available in July 2012 for the creation of 2011 mid-year population estimates, produced annually by the ONS Centre for Demography and calibrated to newly-available census data following each census.

1.1 WHAT IS QUALITY AND HOW IS IT MANAGED IN THE 2011 CENSUS?

The 2011 Census has four strategic aims:

- to provide high quality statistics that meet user needs
- to build confidence in the final results
- to provide value for money solutions, and
- to protect, and be seen to protect, confidential personal census information.

To achieve those aims, a series of critical success factors (CSFs) are being developed which will give tangible evidence of whether those aims have been met. Quality throughout the census operation will be managed using a quality model (defined in the ONS 2011 Census Quality Strategy [1]) that involves:

- design quality
- operational quality management
- quality assurance
- quality measurement and reporting
- the production of high quality population statistics

1 The ONS 2011 Census Quality Strategy is available at: www.gro-scotland.gov.uk/files1/stats/census-quality-strategy-2011.pdf

The Data Quality Assurance Strategy contributes to these objectives by validating census data from the period before 2011 Census day, when management information and early census returns will provide early evidence of response patterns and characteristics, through to the publication of outputs. Raising data quality through field procedures for example is not covered here but is addressed through operational quality management, also referenced in Section 1.4. While it is not intended that this strategy will include the quality assurance of individual outputs, the Data QA Strategy will provide census population and subpopulation counts and variable frequencies against which outputs can be validated.

The challenges for the 2011 Census include:

- the accurate and appropriate counting of increasingly complicated households and living arrangements
- respondent apathy given too much junk mail
- public wariness of joined-up data and their handling in government
- increased levels of internal and international migration
- difficulties in finding people at home [2,3].

The 2011 Census design addresses these issues, drawing on the 2001 Census and international experiences, particularly the censuses of Australia, New Zealand, the US and Canada.

Regardless of the methodology used and ONS's emphasis on best practice throughout census operations, different types of error will affect the census at different stages of the processing. These are likely to include errors or omissions in the address register, respondent error, errors by field staff, coding errors and the compounding of existing error or new errors introduced during data processing. The Data Quality Assurance Strategy will manage data quality through the identification and correction of each type of data error.

There is a fixed time frame between census day and production in the summer of 2012 of 2011 mid-year estimates calibrated to the 2011 Census. This places limitations on the extent of quality assurance activity that is possible, with an inevitable trade-off between accuracy and timeliness. Any delay in data processes poses a risk to timely QA of results. The proposed strategy reflects a considered balance between data relevance, accuracy, timeliness and coherence [4].

The data accuracy that can be achieved reflects the methods and resources in place to identify and control data error and is therefore constrained by the imperative for timely outputs. 'Timeliness' refers to user requirements and the

2 Treasury Select Committee Eleventh Report of 2007-08, Counting the Population, London: The Stationery Office, viewable at: www.publications.parliament.uk/pa/cm/cmtreasy.htm

3 Statistics Commission (2007) Counting on Success. The 2011 Census- Managing the Risks, Report No 36, London: The Statistics Commission.

4 UNECE (2008) 'Census Quality and Evaluation', presented to the Joint UNECE/ Eurostat meeting on Population and Housing Censuses, 13-15 May 2008, Geneva.

guiding imperative for the 2011 Census is to provide census population estimates for rebased 2011 mid-year population estimates in June 2012. 'Coherence' refers to the internal integrity of the data, including consistency through the geographic hierarchy, as well as comparability with external (non-census ONS) and other data sources. This includes conformity to standard concepts, classifications and statistical classifications. The 2011 Data Quality Assurance Strategy will consider and use the best available administrative data sources for validation purposes, as well as census time series data and other ONS sources. A review of these sources will identify their relative strengths and weaknesses. The relevance of 2011 Census data refers to the extent to which they meet user expectations. A key objective of the Data Quality Assurance Strategy is to anticipate and meet user expectations and to be able to justify, empirically, 2011 Census outcomes. To deliver coherent data at acceptable levels of accuracy that meet user requirements and are on time, will demand QA input that is carefully planned and targeted. Mechanisms to achieve this include the identification and prioritisation of:

- key stages in data production
- data items that are critical for census outputs, which involves prioritising the variables used to create census population estimates
- local authority districts posing the biggest challenge or unique challenges to enumerators and where data uncertainties are most likely to be concentrated

For 2011, local authority districts (LADs) [5] will be categorised in advance to support a stratified approach to QA. While data will be delivered to ONS in batches of LADs, called delivery groups, QA will be carried out using LADs as the key analysis unit, though regional, national and cumulative total checks are described later. The most problematic LADs will be identified by reviewing three sources:

- research being undertaken in ONSCD on LADs experiencing high population change
- research by the Census Design Authority on difficulty of enumeration
- research being undertaken by ONS Methodology on estimating response rates in the 2011 Census

It is possible that all approaches will identify the same LADs. For areas that pose a greater challenge, supplementary information will be prepared for consideration by the QA Panel (described below). The prioritisation of LADs will be shared with census regional champions.

In addition, a 'control set' of less complex or challenging LADs will be constructed to assess data quality issues that are not compounded by exceptional enumeration difficulties. This may be used to test and tune tolerances for comparator data.

5 For the purposes of this document, the term LAD includes metropolitan and non-metropolitan districts, London boroughs and unitary authorities

This customised approach to the QA task aims to ensure that the right balance between the competing dimensions of data quality is maintained.

1.2 THE OBJECTIVES OF THE 2011 DATA QUALITY ASSURANCE STRATEGY

The objectives of the 2011 Data Quality Assurance Strategy are:

- to ensure that 2011 Census outputs are fit for purpose and meet user expectations
- to be able to understand differences between census population estimates and rolled-forward mid-year population estimates or other survey and administrative sources and explain these to i) the QA Panel, which has responsibility for recommending approval or rejection of each census population estimates, and to ii) census users through the production of informative metadata.
- to ensure that census information on population structures and characteristics are accurate
- to be transparent in the methods and implementation of data QA
- to plan and implement QA activities in partnership with census stakeholders
- to provide indicators of census quality in quality reports that will be released at the same time as census outputs
- to provide timely census QA input to the production of 2011 mid-year population estimates in 2012.

User expectations will be directly addressed following consultation in 2010 with users at all levels, including statisticians and analysts in LADs. ONS will consult LADs, with the support of the census regional champions and regional statisticians, seeking feedback on local issues, concerns, expectations and observations on the accuracy of the 2010 mid-year population estimates for their areas. This information will help inform and structure QA activity.

1.3 GUIDING PRINCIPLES FOR THE 2011 DATA QUALITY ASSURANCE STRATEGY

The following principles will guide all data quality assurance activity:

- 2011 Census data and management information will be extracted at the earliest opportunity to enable the prompt identification of major systemic and data problems so that they can be addressed and resolved through changes to data capture, processing and statistical adjustment systems
- Every effort will be made to anticipate the questions and challenges that may come from external users and the data QA will allow ONS to target areas of concern

- The biggest user focus will be concerned with local authority population estimates and their QA will be the highest priority
- Subject matter experts will be actively involved at key stages of the data QA
- To enable the QA team to focus on anomalies and ad hoc analysis, QA systems will be pre-programmed and automated in advance to identify likely sources of error/ challenge. Screening checks on the data will be parameter-driven rather than hard-coded, offering the flexibility for non-programmers to adjust them and thereby reduce programmer input during the live QA process.
- Less straightforward and hard to reach groups such as under-ones, migrants, armed forces, people living on caravan sites, in temporary accommodation, in apartments, in gated communities, in communal establishments, sleeping rough and in houses in multiple occupation, will be identified in advance and the risks associated with each assessed. Strategies for evaluating the data quality for each group will then be implemented, based on the risk assessment.
- Spatial analysis techniques are being investigated with a view to using them to help the identification of outliers and data patterns

1.4 SCOPE OF THE DATA QUALITY ASSURANCE STRATEGY

The 2011 Census Data Quality Assurance Strategy is concerned with monitoring and managing the quality of 2011 Census data for England and Wales. Collaboration with the General Register Office for Scotland (GROS) and Northern Ireland Statistics and Research Agency (NISRA) aims to ensure a consistent approach. The Welsh Assembly Government (WAG) is being consulted to ensure its views are represented and reflected in the Strategy.

Management information created throughout 2011 Census operations will report on the progress of census processes and support ONS in ensuring critical success factors are being met. Management information will provide valuable evidence on response rates prior to receipt of data from census forms. Together with the address register it will provide quantitative and qualitative information on enumerators' experiences in the field. The production of management information is described in the ONS Management Information Strategy. ONS is still considering the best way to disseminate management information to local authorities and regions.

Further work is planned to identify roles, responsibilities and strategies for ensuring data errors are corrected in the most effective way. These will be elaborated on in the operational quality management plan.

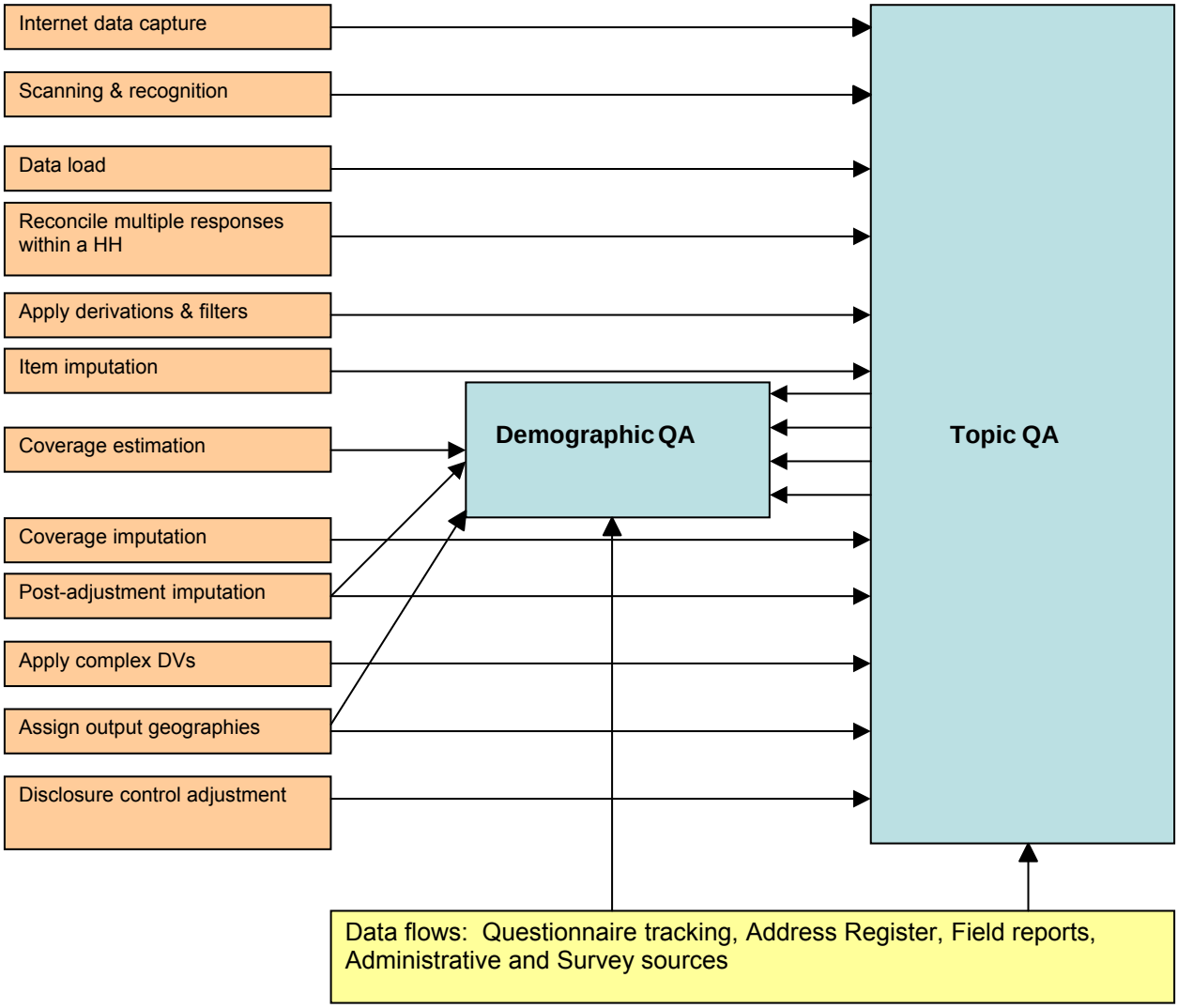
1.5 COMPONENTS OF THE DATA QUALITY ASSURANCE STRATEGY AND THEIR TIMING

Data quality assurance is organised into two related and inter-dependent activities, referred to here as topic QA and demographic QA (see Figure 1). Topic-level QA involves an analysis of the data during and after processing to ensure data items and the outputs for low-level geographies and specified population subgroups have acceptable levels of accuracy. Input from topic experts will ensure that variable distributions at different geographic levels are consistent with the best available alternative sources.

Demographic-level QA ensures that census population estimates and household structures at national, regional and local authority levels pass a series of validation checks, detailed below. Demographic QA overlaps with topic QA because the organising principle for 2011 data processing and delivery is geographic. Data will be supplied to ONS in batches of LADs. LAD counts are the building blocks for regional and national estimates. Both topic and demographic QA involve counting and monitoring the geographic distributions of particular subgroups, for example students, pensioners and those in the armed forces. In addition, both topic and demographic QA are concerned with the integrity of household sizes and structures. Error in the distributions of key demographic variables will skew census populations, as will error in the variables used to estimate coverage such as economic activity and ethnicity. Conversely, error in census population estimation could generate implausible variable distributions. This interdependency demands rigorous cross-checking.

Topic QA checking is concentrated early on in data processing because the opportunity for correcting problems in processing systems decreases over time, giving way to correction through edit and imputation. Once errors arising from field and data capture operations have been addressed and field operations are complete, less checking should be required to maintain quality levels. However it is recognised that some errors may be localised and therefore emerge late in processing or may only emerge after a substantial volume of data has been processed, for example industry and occupation data anomalies may be visible through regional or other aggregate-level geographic comparisons. An indicative timetable of checking activity is summarised in Table 1. Week 1 begins the morning after census night. (For a longer-term timeline of quality assurance activities please refer to Appendix J).

Figure 1 Topic and demographic QA and census data processes



The level of questionnaire completeness for both paper and internet submissions will be calculated after data load and following the reconciliation of multiple responses and the creation of derivations and filters, to provide a net completion rate.

Table 1 Timing of QA activities

Week*		Activity	
-3 to 7	7 March – 8 May 2011	Management information monitoring Examination of preliminary data from supplier system, focusing on web responses which require less processing than paper questionnaires.	
3 to 11	11 April - 12 June 2011	Topic QA: detailed analysis of data items in the preliminary batch may provide early warning of processing problems. Some data quality issues to be addressed through alterations to processing, but most likely through edit, imputation or manual corrections for small population groups.	
12 to 15	13 June – 10 July 2011	Topic QA: intensive checking of regular batches of grouped LAD data	
16 to 20	11 July- 14 August 2011		Demographic QA: LAD analysis and sign-off by QA Panel
21 to 50	15 August 2011- 4 March 2012	Topic QA: quality maintenance checking of regular batches of grouped LA data	
51 to 54	12 March- 8 April 2012	Topic QA: re-checking of post coverage adjustment variable distributions	Demographic QA: QA under- and over count estimates and reconcile LAD, regional and national estimates
55-66	9 April – 1 July 2012	Contingency period**	

* Enumeration, including follow-up, runs from week -3 to week 6 and the census coverage survey takes place between weeks 6 and 10.
**Supplementary analysis that will include linkage of administrative data will begin in April 2010 and the results will inform adjustments made in April-June 2012.

1.6 INTERFACES AND DEPENDENCIES WITH OTHER DATA SYSTEMS

The Data Quality Assurance Strategy will use the data quality monitoring system to inspect and analyse census data using data feeds from a range of ONS and other sources. The precise nature and functional requirements of these interfaces are to be developed (see indicative list in Appendix B).

The Data Quality Assurance Strategy is not designed to replicate any of the QA activity being conducted by ONS census contractors. The ONS quality team is contributing to the development of suppliers' operational management plans and data quality management plans (DQMPs). The complementary nature of activities of the contractors and ONS is illustrated, for example, by the DQMP covering coding activity, which seeks to ensure the consistency

and accuracy of allocated codes. The Data QA Strategy will confirm that the codes meet statistical expectations, using the best available comparator data.

1.7 IMPROVEMENTS IN DATA QA SINCE 2001

The methodology for 2011 QA has explicitly drawn on 2001 experience, through the examination of 2001 policies, systems and publications, a review of the literature following the 2001 census and interviews with those responsible for 2001 QA. The lessons learned have shaped the proposed strategy so it should cope with data quality challenges that can be anticipated because they were present in 2001. Unanticipated challenges, such as a 2011 variant of the foot-and-mouth epidemic, will be the ultimate test of the strategy's effectiveness.

Specific enhancements in 2011 include:

- an enhanced address register against which returns will be monitored
- availability of management information, including 'soft' data from field operations, as qualitative and quantitative evidence to inform the quality of LAD estimates
- explicit and planned prioritisation of QA processes
- planned unit record linkage to address quality concerns prior to data release
- use of newly available administrative datasets for validation purposes
- integration of topic and demographic QA within a single strategy, carried out by a single team

1.8 SKILLS AND ROLES REQUIRED FOR DATA QUALITY ASSURANCE

The 2011 Census quality team will have responsibility for delivering the Data Quality Assurance Strategy, which is summarised and prioritised in Appendix C.

Senior researchers will have responsibility for topic and demographic QA, respectively, with both contributing to the validation of counts and estimates for key population subgroups. Research officers will undertake data analysis, primarily quantitative, to analyse and prepare comparator datasets, agree and implement tolerance levels for expected data values and the analysis and interpretation of findings on rehearsal, census and address register data.

The team will need to be skilled in the use of analytic software including Excel, SAS, Data Visualisation, Spatial Analysis and database applications. The quality assurance task involves the manipulation and reporting on a vast range of datasets, including multiple versions of the same one. It is essential these are correctly and unambiguously configured and that their preparation and use in various software applications is rigorously managed. This work will be supported by a senior executive officer, who will also be fulfilling the same

role in their remaining time to support the maintenance of the 2011 Census internal management information system, which will share the same or similar software applications.

The researchers and senior executive officer will be supported by executive and administrative officers who will make sure that data management, transformations and supporting systems (databases, metadata) operate efficiently.

Skills within the quality team will be developed through a range of training including on software applications, academic courses in methods and substantive areas, coaching within and from beyond the census team, peer collaboration (see collaborating ONS teams and divisions, below) and through international observation and review. Team experience and the 2009 rehearsal will boost data knowledge and handling skills and will build awareness of quality issues in the data.

Good working practices including transparency, particularly in data analytic work, will support continuity in the context of staff departures. The quality team will also be supported by collaboration with other business areas within ONS:

- The ONS Centre for Demography, who will provide consultancy advice and expertise in data sources and methods including QA information drawn from the ONS longitudinal study
- The ONS Methodology Directorate, which has responsibility for coverage estimation. In addition, the Methodology Directorate will provide expert advice, prototyping and coaching on data visualisation and spatial analysis software and will advise on statistical issues including data linkage
- The neighbourhood statistics team, which will provide advice and data on comparator sources and will assist in data analysis and analytic methods
- Administrative Sources and Integration Directorate, which will coordinate requirements for secure legal gateways and make available administrative microdata sources for census data QA
- Analysis directorates which will provide topic expertise for data validation.

Data quality assurance will also benefit from inputs from external stakeholders, discussed further in Section 3. This will include non-ONS topic experts and academics.

2 Achieving Our Quality Objectives

2.1 QUALITY ASSURANCE OF CENSUS TOPICS (TOPIC QA)

2.1.1 The variable checking framework

The principal focus will be on variables that create population and housing counts and those used to support key uses of census demographic information, for example mid-year population estimation and the distribution of resources to local authorities. This prioritisation follows the practice used by Statistics New Zealand [6] and has involved categorising census variables as high, medium or low priority [7]. A lesson from 2001 census was that data quality monitoring needs to be targeted to identify errors that might have a substantial impact on outputs [8,9] rather than getting bogged down on quality issues relating to few records or lower priority variables.

Pre-programmed checks will be made on census data at different geographic levels and at different processing stages. Checks that are actionable will be the highest priority. Automating the checking means that data validation will be faster, allowing more scope for correcting problematic processes (the most likely solution being through edit and imputation) and freeing up resource in the quality team to focus on unexpected or anomalous results. The full list of checks and processes being checked will be agreed in advance. An indicative framework is described in Appendix D, together with an assessment of data availability for checking.

In 2001, a number of 'soft checks' were made on the data, for example numbers of step-children. It is proposed to check the plausibility of these distributions as part of the data QA.

2.1.2 Comparator data and statistical tolerances

At the stages of data processing identified by the checking framework, 2011 variables will be compared with external sources. Comparator datasets will be identified through consultation with ONS specialists and topic experts. Variables known to vary geographically will have additional checks for implausible distributions. An example would be farmers in occupational data for Greater London.

---

6 See the Statistics New Zealand (2008) 2006 Quality Management Strategy (QMS) Summary Report. Available at www.stats.govt.nz/census/about-2006-census/methodology-papers/quality-management-strategy-report.htm

7 'Prioritisation of Census Variables' was approved by the Design and Change Control Board on 3 September 2008.

8 NISRA, '2001 Census: Data Validation', can be viewed at: www.nisranew.nisra.gov.uk/census/censusevaluation/data_validation.pdf

9 ONS (2003) Census 2001 Review and Evaluation, Data Validation: Executive Summary, Hampshire:ONS. Viewable at: www.statistics.gov.uk/census2001/reviewevaluation.asp

The quality team will identify, through consultation with colleagues in the Methodology Division, tolerance levels within which 2011 Census data will be permitted to fall. Those outside of the specified tolerances will be referred for further investigation. A Red-Amber-Green system built into the automated checks will highlight distributions that can be accepted with caution and those that need further checks. It is likely that tolerance levels will vary by geographic level and area, with greater variation and therefore higher tolerance anticipated where areas have experienced rapid recent population change, as signalled by the latest information on population flux from the ONS Centre for Demography. Consideration will be given to adjusting comparator data in areas that have experienced high known levels of population change since the data were originally captured. The QA team will draw on expertise in ONS Methodology to draw up a unified approach to setting tolerance levels within the various census topic areas. Topic experts will approve expected values for variables relating to their subject specialism.

Special attention will be paid to improbable variable combinations arising from multivariate analysis. These have the potential of being highly visible (though containing few cases) which could raise questions about the credibility of the data. Conversely, they could represent real phenomena and their data values should not therefore be suppressed or edited if this is the case, for example persons born outside of the UK could have age-marital status combinations that reflect overseas culture and law. Statistics New Zealand distinguishes between impossible variable combinations, which require correction, and improbable ones, which require monitoring and review, and which it may be appropriate to leave in the data in recognition of and respect for respondents' intended responses. It is proposed to adopt this approach.

2.1.3 Change over time in the topic QA strategy

In recognition that early topic QA processes will identify and correct any systematic errors, following the initial topic QA processes, a reduced framework of checks will ensure that quality is maintained and no new error is introduced. It is envisaged that a subsample of say 1 in 10 LADs in later batches will be as thoroughly checked as the first batches were. The remaining 9 out of 10 LADs will receive adequate checking to ensure that quality standards are being maintained. The latter checks will compare values and distributions at the beginning and end of data processing and check for anomalies. There will be the ability to revert to more comprehensive checking if new systematic errors begin to emerge and to revisit previously checked LADs to verify that a newly discovered error was not previously missed.

Some variables that are difficult to impute will be checked for completeness and plausibility at the end of processing. Examples would be address one year ago, workplace address and address of second residence.

2.1.4 Topic QA at regional and national levels

It is possible that errors within tolerance at LAD level generate unacceptable errors at the regional or national level, suggesting systematic error that the LAD-based approach does not capture. For this reason, a cumulative distribution of key variables will be maintained and compared to an expected cumulative distribution, with stratified tolerances. LADs that take the cumulative distribution close to or outside of tolerance will be highlighted for further analysis by the data quality monitoring system. Likewise, once outside of the range of expected values, cumulative distributions will be closely monitored to determine whether LAD adjustments are required.

2.1.5 The role of topic experts

Specification and details of pre-programmed checks will be developed in conjunction with topic experts, who will most often be ONS staff specialising in the respective subject area, but may involve experts from other government departments. Topic experts will also advise during the QA checking process and may recommend further analysis to assess data fitness for purpose. Topic expertise will also be available from Northern Ireland Statistics Research Agency (NISRA), General Register Office for Scotland (GROS) and the Welsh Assembly Government (WAG). Input from the devolved administrations will have particular relevance for regional, border and migrant characteristics.

2.1.6 Interface with population counts and structures, and population subgroups

The topic QA will monitor and collaborate with the demographic-QA processes on population subgroups including but not limited to:

- babies aged under one
- young men, who had high rates of missingness in 2001
- over-85s
- those serving with UK armed forces
- those serving with foreign armed forces
- international migrants, both long and short-term
- prisoners
- residents and staff in communal establishments
- minority ethnic groups
- same-sex couples
- people in civil partnerships
- private renters
- students
- those with a second address
- non English speakers
- internal migrants

Special attention will be paid to areas where problems of enumeration are anticipated, or where examination of management information from field operations, including the results of address checking, has highlighted particular difficulty in enumeration, for example:

- areas with lots of holiday homes
- areas containing caravan sites
- areas with large numbers of second residences
- areas with high multi-occupancy
- regeneration areas
- areas with low enumeration or high variability

2.1.7 Modal effects: monitoring the difference in paper and internet data

Examination of early internet responses in 2011 and of 2009 rehearsal data will provide an indication of how mode of collection impacts on data quality. Internet data capture (IDC), which is new to the England and Wales census in 2011, offers an opportunity for efficient data collection with enhanced data quality, due to the provision of validation messages that inform a respondent if a particular answer in incomplete or invalid. In addition, 'automated skipping' will route respondents through the questionnaire in such a way that they are not presented with irrelevant questions (such as employment questions for under-16s), thereby reducing respondent burden.

However the modal bias as a result of these innovations is currently unknown. The effect is likely to vary by question and the overall impact on data quality due to differences between paper and IDC respondents will be clearer following the 2009 rehearsal. To unravel these differences, a selection of checks described in the variable checking framework presented in Table E will be repeated for internet responses only and the differences will be analysed. These checks will be in addition to those being carried out by the 2011 Census internet data capture team.

2.2 QUALITY ASSURANCE OF POPULATION COUNTS AND STRUCTURES (DEMOGRAPHIC QA)

Demographic level QA is concerned with ensuring that where census population estimates and structures diverge from non-census sources, including mid-year population estimates, the differences are understood and can either be explained or prompt a review of the census estimates. The resolution of conflicting or inconclusive results in July 2012, when rebased mid-year population estimates are required, is discussed in Section 2.2.12. Quantitative and qualitative methods will be used.

The 2011 Census will build on 2001 methodology to adjust for under- and over-enumeration [10].

Individuals and household records from a large survey, the census coverage survey (CCS), of approximately 500,000 individuals, will be matched against census records. Dual system estimation will estimate census undercount and overcount and further modelling will create estimated populations. Population totals for five-year age groups by sex will be estimated for each local authority district (LAD). Batches of around 300,000 households, 500,000 individuals within LADs will be received from the data supplier. Record imputation will add the population estimated to have been missed to the census database and item imputation will follow, to correct data anomalies arising from the coverage imputation. Timing constraints mean that the CCS sample design has to be set before management information can provide evidence of response rates for different areas. Contingency planning for a reserve survey sample in areas with low response is under consideration.

The demographic QA aims to validate census population estimates following adjustment for coverage at a local, regional or national level. First, the plausibility of the estimated population totals, by sex and five year age group within LADs, will be assessed. A second check following coverage and item imputation, including single year of age, will quality assure the demographic and other key characteristics of the imputed population and reconcile national, regional and LAD populations. Checks at both stages will draw on contemporaneous or adjusted historic external sources, described below. A review of alternative sources will ensure that the most reliable are used as data comparators.

The impact of each LAD estimate and adjusted population subgroup on regional and national estimates will be monitored through cumulative distributions, which will be compared to external sources on an ongoing basis. In addition, a number of population subgroups have been identified that have, in the past, presented particular challenges in enumeration. These will be analysed at LAD level and above to ensure the feasibility of their distributions at different geographic levels.

2.2.1 Quality assuring the estimated population totals through quantitative and qualitative analysis

The objective is to (1) assess whether the estimated populations are consistent with a range of comparator datasets, and (2) decide whether the LAD estimates are ready for referral to the QA Panel, described below, for acceptance or whether coverage estimation needs to be revisited.

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

A number of comparator data sets, including various administrative sources and mid-year population estimates, will be used to create expected values for the population distributions and a number of key indicators. The pre-adjusted, raw census distributions will also be compared against the estimated populations. Key indicators will include, but not be limited to:

- population aged under one
- population by five-year age group up to the age 85+, by sex
- sex ratios by age group
- young and old dependency ratios
- number of households
- household size
- fertility rates
- mortality rates

A check against expected values will be made on the volumes of dummy forms for unoccupied properties. This will also form part of management information monitoring.

This LAD-based analysis will be supplemented with a comparison of population distributions and key demographic indicators for each LAD against a growing cumulative total, at regional or national level. Where the census has picked up a population trend that is not visible in comparator sources, a range of qualitative evidence, described in section 2.2.7, will provide contextual information to help the QA Panel to decide on the next steps, particularly whether the estimates are robust.

At this stage the estimates are by age, broken into five year groupings, and by sex. Following imputation, single years of age are available and these imputed data, which will also at a later stage include adjustment for overcount, will be reviewed again by the QA Panel.

2.2.2 Quality assuring the post coverage imputation Census database

After coverage imputation and corresponding item imputation are applied to the added population, further checks by population subgroup and by single year of age will be made. The same administrative sources as were used to check five year age and sex groups will provide checks for single years of age. This is to ensure the consistency of distributions of ages within and between the age-sex groups that have already been accepted.

2.2.2.1 Population subgroup analysis

Adjusted counts and distributions (LAD, regional and national) of population subgroups will be compared with alternative data sources (including but not restricted to those in Appendix H). Anomalous or inconsistent results will be explored and advice sought from topic experts. The sub-groups to be examined are those monitored by topic QA processes, described in section

2.1.6. There is a substantial overlap between the groups identified here and those identified for prioritisation in field operations [11]. This list may be amended in light of the review of comparative sources. Sources used as comparator datasets in 2001 are listed in Appendix E.

The age structure and characteristics of people for whom English is not their first language will be compared with external sources to check the size and characteristics of this group. This additional attention reflects the high risk of under-enumeration of this group. A stratified approach will be taken to the QA of in-migrants' data (based on the intention to stay question). Migrants included in the main output base will undergo more extensive checking than short-term migrants.

2.2.2.2 Cohort analysis

The plausibility of both the adjusted estimates and the post-imputation database will be checked through cohort analysis. This will involve identifying 2001 cohorts in 2011 data and comparing their characteristics over time. Likewise 2011 cohorts will be compared with aged-on cohorts from mid-year population estimates.

The plausibility of student ratios within cohorts will be compared with 2001 patterns and external sources. Age structures within communal establishments and residents' characteristics that vary by age will be compared with 2001 patterns.

2.2.2.3 The address register as a QA source

An important methodological enhancement in 2011 is the development of an authoritative and up to date address register coupled with a questionnaire tracking system. These will be critical for informing and guiding effort in the field, especially follow-up, but also for assessing the quality of the enumeration. An accurate address register and effective questionnaire tracking systems are the cornerstone for successful enumeration and population adjustment. The address register will be created by the integration of the Royal Mail postcode address file, the Ordnance Survey Master Map Address Layer 2 and the National Land and Property Gazetteer. Discrepancies between these primary sources will be referred to local authorities for resolution. Prior to census day, checks will be made in areas with the highest level of discrepancy between the three sources and in areas with high levels of multi-occupation.

The address register provides the household frame for tracking questionnaires. It will be updated using evidence gained during the operation, for example from the contact centre, during follow-up or from post-back.

11 ONS (2008) Hard to count group matrix. Information and Guidance paper- management summary, paper presented to the 2011 Census Programme Board, July 2008.

Addresses could be 'deactivated' because they are found to be non-residential, derelict, demolished or duplicate entries, or they could be created if newly found. Management information indicators will monitor volumes of deactivated addresses.

During follow-up, 'dummy' forms will be created for addresses where no questionnaire has been returned. The dummy form will identify addresses where the household is absent, where no questionnaire was returned and contact could not be made, that are second or holiday addresses, are vacant or where householders have refused to supply census information. Corresponding categories for communal establishments are absent establishments, those that did not return questionnaires and where contact could not be made, vacant communal establishments or those that refused to return questionnaires. Dummy form information will be received at ONS alongside census data within respective data batches.

It is the intention that every address in the household frame should be accounted for and response rates calculated and monitored. For each LAD, a reconciliation exercise will identify and investigate gaps. Dummy form information is primarily collected to inform the placement of missed households in the imputation process, but it also supports use of the address register for QA purposes:

- The coverage survey provides an opportunity to check the accuracy of address register and dummy form information for households and for communal establishments with up to 100 bed spaces. Administrative sources can be used to check the larger communal establishments
- To investigate challenges to LAD estimates based on external sources, outlined in section 2.2.12 below
- Address-based validation checks for selected LADs. For example, council tax registers containing information on types of discount may be available at LAD level. Aggregated information on the incidence of single-person households, households with disabled residents, students and nurses from council tax records can be compared with the address register and census results to check topic data quality, imputation quality and coverage.

Conversely, the address register can be used to assess and validate coverage estimation. For example, where coverage estimation implies more addresses than there are on the address register it may be necessary to check data and assumptions.

Consistency in the definition of output area geography between 2001 and 2011 will aid comparisons at this low geographic level. In 2001, census population estimates, resulting from the coverage adjustment, were applied at the LAD level and not controlled at lower geographies. A decision on the 2011 approach to lower super output area (LSOA)-level adjustments will be made during 2009. There is scope for large differences in 2001/2011 populations at LSOA level that are due to imputation and which could undermine public and stakeholder confidence in the 2011 estimates. To

address this and to focus QA resource, 2001/2011 absolute population differences will be calculated and ranked and a sample of those areas with the largest difference attributable to imputation (in either 2001 or 2011) will be identified. All LADs where large differences cannot be explained by imputation will be investigated, drawing on all available qualitative and quantitative evidence. The aim will be to understand how the differences arose and to thus justify or adjust 2011 estimates.

2.2.3 Reconciliation of national and local estimates

As well as reviewing and formally accepting LAD estimates, the impact of changes at the sub-national level on regional and national totals will be monitored.

The QA Panel (described in Section 3) will consider the cumulative totals for all approved areas, which will give an early indication of adjusted regional and national profiles. This will comprise:

- cumulative population distributions for the approved areas against national distributions from comparative datasets, such as mid-year population estimates
- cumulative population totals for the approved areas against cumulative expected values derived from comparator datasets.

The summary indicators, these cumulative distributions and their national and cumulative comparator equivalents will include but not be limited to those identified at the first stage of demographic QA and described in section 2.2.1.

A list showing which areas are included in the cumulative totals will accompany national graphs. This will enable the panel to assess whether the areas processed and agreed are skewed towards particular population characteristics. Repeatedly comparing aggregated sub-national distributions with national profiles will give an early warning of where adjustments appear to be leading to under/overcount at the national level.

Some QA will be undertaken at regional levels. This will occur where an alternative data source provides a good comparative source at higher levels of geography but is unavailable, unreliable or incomparable at LAD level. For example, survey data such as the International Passenger Survey are only robust at higher geographies. Additionally, characteristically different areas such as Inner London will merit separate consideration. ONS's regional statisticians and local experts will be invited to contribute evidence for the regional-level QA (described more fully in section 3).

A reconciliation of LAD, regional and national estimates will be carried out after the final adjustment for overcount and usual residence, as described below. Following the 2001 census, evidence from the ONS longitudinal study was used to assess whether census adjustments for under- and over-count explained the absence from the census of LS members who were not known

to have died or left England and Wales [12]. This analysis included an adjustment for longitudinal attrition. The intention for 2011 is to synchronise LS linkage and analysis so that this longitudinal evidence will be available as an external check to support census demographic QA. The LS sample size (1 per cent) means that evidence at LAD-level will not be reliable, but could be available at higher aggregations.

Monitoring of population subgroups at regional and national levels is covered in section 2.2.2.1.

2.2.4 Evidence of multiple enumeration and adjustments for usual residence

In the 2001 Census the main reasons for people being counted more than once were:

- they were students counted at both parents' and term-time addresses
- they were counted in communal establishments and in the family home
- they were counted at two addresses, one of which was also listed as their 'address one year ago'. These were typically people moving house around census day who completed forms at both addresses
- they were children of separated parents counted with both parents or with both a parent and another relative
- they were people with more than one property counted at both
- they were young adults counted with friends/ flatmates and parents
- they were people who filled in more than one form at the same address or who entered themselves more than once on a single form.

In 2011, multiple enumerations within an address may also arise due to completion of both paper and internet forms. Addresses will be linked via the questionnaire tracking system and within-address multiple enumerations in 2011 will be identified through name matching in downstream processing. The primary record will be flagged for inclusion in the appropriate population base.

There will not be enough time to attempt full reconciliation of records for visitors/ second addresses/ multiple enumeration in the census so the QA strategy will include creation of a matrix showing where people could be over-counted or counted at the wrong address and will identify the sources of information that are available to quantify and geographically locate each category. This information will help to assess whether the difference between raw counts and final estimates are plausible. Appendix F lists some of the categories to be identified, based on 1991 data on visitors, usual and absent residents in the ONS longitudinal study (1991 data are more relevant than 2001 in this context, given similarities in 1991 and 2011 enumeration bases).

12 Blackwell, L., Lynch, K., Smith, J. and Goldblatt, P. (2003) Longitudinal Study 1971-2001: Completeness of Census Linkage, London: ONS. Also found at: http://statistics.gov.uk/downloads,theme_population/LS_no10.pdf

This work will be informed by routine work carried out within the Centre for Demography to identify, define and estimate population subgroups that impact on mid-year population estimates and the calculation of different population bases.

Evidence to be used for the 2011 estimation of overcount is likely to include:

- results of matching a sample of visitor and usual residence records
- results of matching second address information
- geographic distributions of visitors' usual residence information to support mapping visitors back to their usual residence
- evidence of missing or misclassified residents/visitors derived from the coverage survey and the quality survey
- information on time spent at second residences drawn from a large sample survey such as the Integrated Household Survey
- the results of definitive record matching in the ONS longitudinal study 1991 and 2011 Census/LS link.
- external information on migration patterns and characteristics of people living outside of the UK for more than twelve months.

These will help to inform the creation of population counts for different population bases, for example one based on where people spend the majority of their time, by improving the QA in areas where there are substantial numbers of second homes.

2.2.5 Local authority strata and the checking framework

In 2001, a 'QA pack' for each LAD provided the information needed for the QA Panel to accept or request further analysis of the census population estimates for that area. For 2011, LADs will be categorised in advance to support a stratified approach to the demographic QA. LADs will be categorised, ahead of census and using the latest available data, in terms of difficulty of enumeration and/ or high recent population change. More extensive checks will be applied to more problematic areas. Time for the QA Panel to assess different types of LAD will be planned according to their complexity.

A table of supplementary information listing key summary data at lower super output area level will be included in the published QA pack for each LAD. This information will be used to identify whether observed under/overcount for an LAD is focussed within a particular output area and to support QA Panel assessments of robustness.

2.2.6 Key data sources and statistical tolerances

2.2.6.1 Assessing quality of comparator data

A review of available administrative data will assess each dataset's strengths and weaknesses and their suitability as an independent source of population

comparison, using a checklist of quality measures. The checklist will be agreed with stakeholders. It will also identify where adjustments are needed for comparability due to definitional differences, timing and timeliness of the data and under/over coverage. Comparative data will predominantly be used at an aggregate level, though microdata (where available) will be used to explore data quality issues and investigate outliers.

The review will draw on expertise in the Improving Migration and Population Statistics (IMPS) work programme, in the ONS neighbourhood statistics team and the Administrative Sources and Integration Division.

The quality team will consider data available to local authorities and devolved administrations as comparators. Special attention will be paid to issues of coverage, for example, DCSF Schools Census data (listed as a potential comparator data set in Appendix H) applies to England only; the Welsh Assembly Government runs its own schools census.

Data used as comparators in 2001 are listed in Appendices E and G. For 2011, new and emerging administrative data sources will be identified and investigated. This will include understanding the measures and checks undertaken by the source owners to ensure the quality of their data and will involve engagement across the Government Statistical Service (GSS). Potential new sources are listed in Appendix H. A decision on comparators to be used will be made in early 2010.

It may be possible to use management information to adjust comparators, where appropriate. For example, dummy forms will be returned for deactivated addresses and, subject to quality checking, these can be used to adjust expected numbers of households that are based on the household frame. Likewise increased knowledge and understanding of flux patterns that result from the address register development, including extensive address checking, will inform the likely reliability of household-based comparators.

2.2.6.2 Methodological enquiries to inform the use of comparator data

The methodology, analysis and assumptions of each comparator dataset will be explored as part of the review detailed in section 2.2.6.1. The commissioning of new analyses will be considered if they will improve the utility of comparator datasets. Methodological investigations could include:

- Linkage of the 2001 CCS and the ONS longitudinal study. This would provide qualitative information on the overlaps and gaps between the CCS, census and health authority registration data (using the LS). Subsequent information on health authority de-registrations can provide proxy information on the likelihood that those found in neither the LS nor the CCS had in fact embarked. It could also provide partial adjustment factors for list inflation in health authority registration data

- Further comparison of LS and administrative data. The fact that the LS is a representative sample of census records coupled with its inclusion of health authority registration data and a history of high-quality, definitive matching between censuses and vital events make it a unique and valuable laboratory for data validation [13] . Further linkage of other administrative sources to the LS would reveal issues of data quality Unlinked administrative sources may also shed new light on areas of LS under-coverage.

- Triangulation of other administrative sources to gain understanding of the timeliness of all three sources and possibly adjustment factors to be applied for list inflation

The census quality team will maintain a watching brief on data linkage studies being carried out in ONS between now and the census, in case they can be used or extended for QA purposes. It is envisaged that pilot linkage studies will be identified and implemented by the quality team around the time of the 2009 rehearsal and subsequently during 2011 to inform census estimate validation. Where administrative data are used in coverage estimation (triple system estimation) they will not be used at the level of aggregation used in QA. Using the same data to check estimates to which they directly contributed creates circularity which would undermine the credibility of the QA process. Linkage of microdata will only be considered where necessary and will be supported by the appropriate legal gateways, including parliamentary data sharing orders under the Statistics and Registration Services Act (2007).

Where high match rates and data quality are necessary, a combination of automated and clerical matching will be used. Limitations on resources mean that clerical resolution will be used sparingly.

2.2.6.3 Methodological enquiries around particular population subgroups

Alongside planning for 2011 QA, several analyses and literature reviews will help to inform the QA strategy and interpretation of findings. Examples include:

- unravelling reasons for the shortfall in young men. Colleagues in the Australian Bureau of Statistics have been investigating this and the results will have direct relevance for England and Wales
- the nature of the relationship between mid-year population estimates and population counts, in particular considering work by critics who argue that the mid-year estimates are overestimates.
- cohort sex ratios and trends over time

13 See Jones, P. and Elias, P. (2006) 'Administrative data as a research resource: a selective audit', ESRC for an assessment of the relative strengths and potential of the LS for administrative data linkage.

2.2.6.4 Statistical tolerance for LADs

As in 2001, expected values for age-sex distributions will be calculated based on comparator data available. The range of expected values for any age-sex group was defined in 2001 as:

'The midpoint of the range of the different comparators plus or minus the range itself'

This definition recognises that although the comparators provide a good estimate of the indicator in question, the comparators themselves contain error and variance. It is likely that the 2011 expected values will use maximum and minimum differences from the midpoint again. Further methodological work, in conjunction with Methodology Division, will establish a coherent approach to tolerance levels for census QA.

The census population estimates with their 95% confidence intervals will be compared with the expected values. Where an estimate and its interval fall wholly within an expected range, this will indicate strong evidence to accept the estimate. However, where an estimate falls outside, further investigation is required. In both cases, the full range of supporting information will be considered.

2.2.7 Supporting evidence

Consideration will be given to a range of supporting information. Listed below are examples of contextual data which will help members of the QA Panel to assess the robustness of census estimates:

- information on the strengths and weaknesses of the comparator datasets identified from the review outlined above
- key management information for an LAD relating to progress and quality of the census and CCS operations from delivery, response, through to processing of data. All management information will be presented against expected values and tolerances.
- edit and imputation reports
- information from census and CCS field staff debriefing sessions and incident recording logs
- information from field staff record books
- a summary for the area based on enumeration intelligence.
- in addition to the above, correspondence between local authorities and ONS, for example on accuracy of mid-year population estimates, the address register or identification of communal establishments.

2.2.8 The relationship between demographic QA and the census coverage adjustment

To ensure the independence of the QA process, administrative data used for dual system estimation will not be used to QA, except possibly when used as a secondary source at a higher level of aggregation.

A lesson learned from 2001 was that the robustness of the methods used could have been supported by publication of the LAD characteristics of non-respondents estimated by the coverage survey and imputed into census records [14].

It is proposed that a running check will be maintained of imputed people and their household characteristics at both LAD and national levels, to monitor the plausibility of the adjustments and for publication to coincide with release of census data.

2.2.9 Quality assurance and the 2011 Census quality survey

The 2011 Census quality survey will involve re-interviewing census respondents. Interviews will be guided by census responses so that inconsistent answers can be probed. This will provide valuable information on the quality of data collected and will highlight areas of the questionnaire respondents found challenging. It may be possible to factor the results of the quality survey into item and coverage imputation processes, subject to scheduling constraints, but the sample size (approximately 2,000 achieved interviews) will not be large enough to support detailed adjustments.

The quality survey aims to interview as many household members as can be found. This will provide data from which the extent of proxy response error can be assessed. These findings will help inform the topic QA processes.

The quality survey will be used to probe issues identified in planning of the topic and demographic QA. An example could be the completeness and quality of information supplied on visitors to households and communal establishments. In addition, data for short-term and long-term migrants could be investigated.

2.2.10 Stakeholder input

To ensure a successful QA process, the knowledge and expertise of a wide range of stakeholders will be sought on the proposed QA methodology and comparator data sources. ONS will be transparent in its approach to give stakeholders the opportunity to comment on and develop the methodology in advance of the QA exercise. Details are described separately in the Quality Assurance Stakeholder Communications Strategy.

14 Simpson, L., Hobcraft, J. and King, D. (2003) 'The 2001 One Number Census and its Quality Assurance', London: Local Government Association, can be viewed at: www.lga.gov.uk/Documents/Publication/onenumbercensus.pdf.

Stakeholders include the ONS Centre for Demography, ONS Methodology Division, Neighbourhood Statistics, the census coverage survey, local authorities and their representatives including the Local Government Association, BURISA, health authorities, census advisory groups, other government departments, GROS, NISRA and WAG.

2.2.11 Outcomes of the QA Panel

There can be a number of outcomes from Quality Panel's review of LAD estimates, described below. Some LADs will go through processes B and/or C more than once.

A. ONS estimates agreed
   - Begin update and imputation processes
   - LAD totals to be included in cumulative national figures

B. Further evidence needed to accept estimates
   - QA team to undertake further analysis as advised by the QA Panel. Advice and guidance to be sought accordingly. Further analysis could include:
     - reconciliation between the address register/ household frame and administrative sources, involving address matching
     - further detailed investigation of administrative sources for specified areas/ data discrepancies
     - sensitivity analysis of data comparisons employing varying assumptions
     - seeking further external sources to reconcile conflicting evidence
     - extending visitor matching sample to a full visitor match
     - using ONS LS to provide longitudinal evidence
   - Additional information to be presented at a subsequent QA Panel meeting. Estimates to be accepted (A) or rejected (B or C)

C. Adjusted estimates rejected
   - Census coverage and adjustment team implement a strategy for amending the coverage estimates
   - QA team to rerun QA analysis on revised estimates
   - Revised QA packs are compiled and presented at subsequent QA Panel. Estimates to be accepted (A) or rejected (B or C)

If estimates are rejected at a regional or national level, all affected LAD estimates will be re-run through the QA process. This may result in the review of LAD estimates that had previously been accepted.

2.2.12 Where agreement of estimates cannot be achieved: supplementary analysis

A key lesson learned from the 2001 census was that the possibility of inconclusive results should form part of census planning. To ensure that supplementary analysis is better planned and resourced, there needs to be a clear set of processes and sources to be used to reconcile conflicting evidence. The strategies to be deployed will be continually reviewed, subject to the availability of new data and information gaps that emerge through planning and implementing the census. The proposed strategy is likely to include a combination of record-level data matching, both automatic (definitive and probabilistic) with the possibility of clerical matching of residual records not matched automatically.

Planning will seek to ensure that supplementary analysis will not jeopardise the availability of finalised census counts for the production of 2011 mid-year population estimates in July 2012. Should new evidence emerge after the release of 2011 Census populations in June 2012, this will be used to adjust and correct subsequent mid-year population estimates rather than prompting a revision to census estimates. Supplementary analysis will begin in April 2010 and will continue through to March 2013, addressing information gaps as they emerge and using new administrative sources if appropriate. This analysis will be subject to approval and guidance from the quality assurance subgroup of the UK Design Methodology Advisory Group, and subsequently from the QA Panel (both described below).

3 Transparency, Peer Review and User Communication

ONS is committed to establishing a consultation mechanism ensuring the development of the QA methodology is a transparent process and users and stakeholders have the opportunity to contribute. Peers and users will be engaged through:

3.1 QUALITY ASSURANCE PANEL

The QA Panel has the responsibility of assessing the quality of the census estimates after coverage adjustment. The panel will review the quantitative and supporting evidence prepared by the QA team to determine whether estimates should be accepted or rejected.

The panel will convene to review QA data from July 2011 but their input and involvement will be critical before this for consultation on the strategy and to agree the contents of the QA packs and the roles and responsibilities of the panel. QA Panel members will meet for the first time in July 2010 to agree the terms of reference and to review the QA approach, in particular the proposed quality information and indicators that will support QA Panel decisions. The panel will then meet quarterly to review comparator data and to advise on further analysis to be done prior to census data QA. The QA Panel will be briefed on census methodology, strengths, weaknesses and likely sources of error.

Table 3 sets out the key milestones that will structure QA Panel activity. It is based on current (February 2009) planning of data processing and assumes that data will be delivered by the contractor as scheduled.

Table 2 QA Panel activity over time

Key dates	QA Panel activity
July 2010-July 2011	Quarterly meetings to agree terms of reference and review comparator data and QA methods. To advise on supplementary analyses.
18 July 2011- 11 Feb 2012	Weekly meetings to agree LAD estimates
12 Feb 2012- 11 March 2012	Adjustments for overcount, visitor and second residence reconciliation agreed
12 March 2012- 27 May 2012	Finalised LAD, regional and national estimates signed off

Between mid-July 2011 and early February 2012 the QA Panel will agree the estimates for 348 LADs (this number includes known changes coming into effect on 1 April 2009), 22 of which are Welsh.

To ensure that quality team resource and QA Panel time is directed efficiently at the LADs which pose the biggest risk in terms of enumeration and data

quality, the QA Panel will convene a subgroup, to meet weekly, which will focus on prioritised LADs. This 'priority QA' subgroup will report on a weekly basis to the main QA Panel, advising in particular on cross-cutting issues or themes that emerge from their review of the LAD estimates. A working assumption is that 25 per cent of LADs will be considered for priority QA. Thus approximately 88 LADs will be assessed by the priority subgroup over a 28 week period, requiring an assessment rate of approximately three LADs each week. The priority QA subgroup will make recommendations to the QA Panel which will need to sign off on average 12 LADs, including priority LADs, each week.

The benefits of having a subgroup of the QA Panel to assess prioritised LADs are:

- The subgroup members will develop expertise in evaluating 2011 data quality issues that are common across the problematic LADs, for example transient population subgroups, large numbers of students or concentrations of second addresses
- The subgroup will commission the quality team to conduct further analysis that it identifies to resolve anomalies. This could include, for example, record-level reconciliation between the address register and selected administrative sources for a particular LAD. The results of such an exercise may be generalisable across several priority LADs. The priority subgroup will ensure that quality team resources are channelled into those analyses that will deliver substantial data quality benefits rather than thorny issues that will be time-consuming to resolve but have limited substantive impact on the estimates
- The subgroup can call upon the insights of local experts to fill information gaps and input additional data

The involvement of local expertise in the QA Panel assessment is critical for building trust and credibility in the census data and methods and for making sure that the estimates are as accurate as they can be. However it is equally important that census data are not influenced by the vested interests of local authorities. Local experts involved in the QA process will be focused on the geographic areas that they are experts in.

Local experts will be identified by ONS regional statisticians in conjunction with the Local Government Association. Options to involve census regional champions are being considered. Up to two experts in each region will be identified and these experts, together with the regional statistician, will be invited to provide quantitative and qualitative information to inform QA Panel decision-making. In addition, the QA Panel will call upon the expertise of census regional champions, their advisors, regional managers and field managers for additional supporting evidence.

It will be essential that momentum in approving LAD estimates is maintained because of dependencies in later data processes, for example post coverage adjustment item imputation. Once all of the LADs have been adjusted for under-enumeration, further adjustment for overcount will be

considered. Adjustments based on visitor and second residence matching will also be considered by the QA Panel at national, regional and LAD levels. This will take place over a four week period between early February and March 2012. Thus every LAD estimate will be considered twice by the QA Panel; in the first instance to approve the five-year age/ sex distributions and secondly following any adjustments for overcount and to approve single year of age distributions and imputation.

Between early March and the end of May 2012, the QA Panel will consider a range of quality indicators at all geographic levels, as set out in this strategy. These will include data item and population subgroup distributions and a number of demographic measures. The plausibility of LAD estimates in these contexts will be reviewed and final adjustments or a review of coverage imputation for particular areas may be considered.

For QA purposes, LADs in Wales will be considered together and Welsh local experts will contribute to the QA process in the same way as regional statisticians and local experts in England.

Membership of the QA Panel will be:

- Census quality team representatives
- Methodology Division representatives (experts in coverage adjustment)
- ONSCD representatives (experts in mid-year population estimation and projection)
- ONS Geography representatives (to advise on the address register)
- Independent demographers, including non-UK members. This will include a member of the LGA.
- Representatives from General Register Office for Scotland (GROS) and the Welsh Assembly Government (WAG).

In addition, the QA Panel may consult field managers to discuss any suspicious data distributions in a particular region.

Meetings will be chaired by the Head of the Census Design Authority.

ONS senior management will review all estimates recommended for acceptance by the QA Panel prior to their release. This will be a formal scrutiny and sign-off process and will involve director and other senior management input from 2011 Census, Methodology, and the Centre for Demography, Analytic Directorates and the National Statistician.

3.2 PEER REVIEW

3.2.1 Quality Assurance Working Group (QAWG)

The Quality Assurance Working Group provides advice, quality reviews and steers 2011 QA activity. The group meets a minimum of twice yearly and includes key representatives from the UK census offices: ONS, Northern

Ireland Statistics Research Agency (NISRA), General Register Office for Scotland (GROS) and the Welsh Assembly Government (WAG). This representation ensures a harmonised UK approach to census QA.

There are ONS representatives across a number of departments to ensure the QA work is integrated with other work streams and that the work, experience and expertise across the office are shared.

The QAWG gains its authority from the Design & Change Control Board (DCBB) in ONS, the Census Programme Board in NISRA and the Census Programme Board in GROS. It must assure these boards that the QA strategy and the proposed methodologies are sound and it raises key decisions to them for approval.

3.2.2. UK Census Design Methodology Advisory Committee (UKCDMAC)

The UKCDMAC was established in December 2004 to review and advise on a range of statistical and methodological issues relating to 2011 Census design. The aim of the group is to achieve high quality, consistent and comparable UK-wide outputs.

The committee includes internal members who represent the UK census offices and external advisors. There are currently external members representing:-

- ESRC
- academia
- GLA
- a local council

The QAWG will seek advice from UKCDMAC on the strategy, proposed methodologies and design decisions relating to the quality assurance of census estimates. This will ensure that QA work will receive scrutiny and input from the wide perspectives covered by this group. A subgroup of the UKCDMAC will be convened to provide ongoing specialist input to the Data Quality Assurance Strategy. In addition, a spatial analysis subgroup will be convened to advise on the utility and application of spatial analysis techniques for the QA task.

3.2.3 International peer review

An international peer review will ensure that the 2011 QA strategy benefits from the latest thinking from census teams overseas.

A presentation of the QA plans was given at the European Conference on Quality in Official Statistics in Rome in July 2008. Other opportunities to engage international colleagues will be exploited where available. Input will be sought directly from countries where the research and thinking on the QA

approach is well developed. This strategy has benefited from the peer review of statisticians in the Statistical Offices of Northern Ireland, Ireland, Scotland, Wales, the United States, New Zealand, Canada, France and Portugal.

3.2.4 Harmonisation across the UK

This strategy seeks a harmonised approach to QA across the UK, and GROS, NISRA and WAG representatives attend various peer review groups through which the strategy has been developed. It has been considered and approved by the UK Census Committee.

3.3 USER COMMUNICATION STRATEGY

As population estimates will receive some considerable scrutiny, it is important that census users are engaged in the planning process. In addition to the user engagement described above, there will be a number of opportunities for users to understand and contribute to QA methods:

- At census roadshows (the first held in Autumn 2008)
- Via established partnership groups. For example, the Central and Local Information Partnership (CLIP) census subgroup and census advisory groups
- ONS published a series of communications to explain the methods being developed for the 2011 Census
- Information and documentation on the data quality assurance strategy will be published on the National Statistics website
- 'Walkthroughs' of the Data Quality Assurance Strategy will be held, at which attendees will be invited to comment on the proposed methodology.

User communication will be ongoing and will culminate in the publication of QA packs for each LAD from mid-2012. These will include the comparative information used by the QA Panel to accept estimates and the key decisions reached in each case.

3.4 OUTPUTS FROM THE DATA QUALITY ASSURANCE PROCESS

Several publications will be produced by the QA team to document the QA process and share the results of QA activity. These reports will include:

- QA packs for each LAD for the QA Panel, to inform its decision-making
- QA packs for each LAD for web publication
- a data quality report, for web dissemination alongside the release of census data and including quality indicators
- a quality report to meet Eurostat metadata requirements
- a quality report which will form part of the Statistics Board assessment process, prior to live running.

4 Data QA Strategy for the 2009 Rehearsal

In October 2009 there will be an England and Wales census field and upstream rehearsal (Northern Ireland will be carrying out a parallel rehearsal at the same time). The rehearsal areas will be Anglesey (34,000 households), Lancaster (62,000 households) and Newham (40,000 households). Together, these diverse LAD types provide an opportunity to confirm the viability of ONS field operations and associated systems and the data processing systems, both within ONS and those provided by external contractors. The voluntary nature of the rehearsal means response rates will be far lower than at census. For this reason, and because there is no coverage adjustment envisaged for rehearsal, aspects of data quality assurance that relate to population estimates will not come into play in 2009. However, this will be an opportunity to:

- test and confirm the utility and effectiveness of pre-programmed data validation checks
- practice and develop ad-hoc analytic skills to supplement automated checks
- test the quality of comparator datasets against rehearsal and address register data and against administrative sources provided by the Rehearsal local authorities (and explore triangulating between these three sets of data)
- test and develop the data quality monitoring system and the associated software used in QA.
- test and evaluate the potential for early QA analyses to inform field activities e.g. targeted publicity to lower responding population subgroups
- review and evaluate procedures for the co-ordinated acceptance of data batches following quality assurance and data validation/consistency checks
- enable the quality team to develop effective working relationships with stakeholders including LADs and topic experts.

Details of QA activities during the 2009 rehearsal are described in Appendix I.

5 Next Steps

Activity	Delivery date
Develop unified approach to setting tolerance levels	Apr 09
Develop supplementary analysis plan	Jun 09
Decision on what happens if QA process is not complete by July 2012- incorporate into revisions policy	June 09
List and QA demographic QA comparators	Sep 09
Policy decision on use of microdata for supplementary analysis and comparators	Sep 09

6 Bibliography

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

Baffour, B. and Valente, P. (2008) 'Census Quality Evaluation: Considerations from an international perspective'. Paper presented at Joint UNECE/Eurostat Meeting on Population and Housing Censuses, Geneva, May 2008. Available at www.unece.org/stats/documents/ece/ces/ge.41/2008/sp.4.e.pdf

Blackwell, L., Lynch, K., Smith, J. and Goldblatt, P. (2003) Longitudinal Study 1971-2001: Completeness of Census Linkage, London: ONS. Available at: http://statistics.gov.uk/downloads,theme_population/LS_no10.pdf

Jones, P. and Elias, P. (2006) 'Administrative data as a research resource: a selective audit', ESRC, National Data Strategy.

Judson, D.H. (2007) 'Information integration for constructing social statistics: history, theory, and ideas towards a research programme'. Journal of the Royal Statistical Society: Series A (Statistics in Society), 170 (2), 483-501.

Martin D.J. (2007) Editorial: 'Census present and future', Journal of the Royal Statistical Society: Series A (Statistics in Society), 170 (2), 263-266.

McBeth, N., McGill, C., Moore, T. and McGregor, D. (2003) 'The Quality Management Strategy (QMS) in the New Zealand 2001 Census of Population and Dwellings and how this shapes the future', paper presented to the American Statistical Association 2003 Joint Statistical Meeting- Section on Survey Research Methods, available at: www.amstat.org/sections/SRMS/Proceedings/

NISRA [ca 2004] 2001 Census: Data Validation. Available at www.nisranew.nisra.gov.uk/census/censusevaluation/data_validation.pdf

ONS (2000) 2001 Hard to count index. One Number Census Steering Committee Paper 00/15. Available at www.statistics.gov.uk/Census2001/pdfs/sc0015.pdf

ONS (2001) A Quality Assurance and Contingency Strategy for the One Number Census. Available at www.statistics.gov.uk/census2001/pdfs/oncinfopaper.pdf

ONS [ca 2002] One Number Census Illustrative Quality Assurance Pack. Available at www.statistics.gov.uk/census2001/pdfs/onc_qa_pack.pdf

ONS (2002) Census 2001 Review and Evaluation - Census Coverage Survey: Evaluation Report.

ONS (2003a)ONS Census 2001 -Dependence within the 2001 One Number Census project. Available at www.statistics.gov.uk/census2001/pdfs/dependency.pdf

ONS (2003b) ONS Census 2001 -One Number Census Quality Assurance information: Quality Assurance themes. Available at www.statistics.gov.uk/census2001/pdfs/borrowing_str_babies.pdf /www.statistics.gov.uk/census2001/pdfs/students.pdf

www.statistics.gov.uk/census2001/pdfs/prisons.pdf
www.statistics.gov.uk/census2001/pdfs/post_stratification.pdf
www.statistics.gov.uk/census2001/pdfs/armed_forces.pdf
www.statistics.gov.uk/census2001/pdfs/foreign_armed_forces.pdf
www.statistics.gov.uk/census2001/pdfs/collapsing_strata.pdf
www.statistics.gov.uk/census2001/pdfs/borrowing_strength.pdf
www.statistics.gov.uk/Census2001/pdfs/1991_underenumeration.pdf
www.statistics.gov.uk/census2001/pdfs/babies_ethnic_pop.pdf

ONS (2003c)Discussion paper: Proposals for an Integrated Population Statistics System. London, Office for National Statistics.
www.statistics.gov.uk/downloads/theme_population/ipss.pdf

ONS (2006a) Enumeration Targeting categorisation to be used in the 2007 Census test. Information Paper. Available at
www.statistics.gov.uk/census/pdfs/EnumerationTargetingCategorisation.pdf

ONS (2006b) The One Number Census – an estimate of the whole population. ONS.
www.statistics.gov.uk/census2001/onc.asp

ONS (2007a) A Review of the Potential Use of Administrative Sources in the Estimation of Population Statistics. Available at
www.statistics.gov.uk/about/data/methodology/specific/population/future/imps/updates/downloads/admin.pdf

ONS (2007b) Guidelines for measuring statistical quality. Available at
www.statistics.gov.uk/downloads/theme_other/Guidelines_Subject.pdf

ONS (2008a) Improvements to Migration and Population Statistics: May 2008 Progress Report. Available at
http://www.statistics.gov.uk/about/data/methodology/specific/population/future/imps/updates/downloads/20_May_IMPS_Progress.pdf

ONS (2008b) 2007 Local Authority Case Studies: Final Report. Available at
www.statistics.gov.uk/about/data/methodology/specific/population/future/imps/updates/downloads/LAreport.pdf

ONS (2008c) Reconciliation of ONS estimates: Comparisons of combined IPS (long and short term migration) estimates with administrative data sources. Available at
www.statistics.gov.uk/about/data/methodology/specific/population/future/imps/updates/downloads/Reconciliation_Exercise.pdf

Redfern, P. (2003) 'Estimating Census Undercount by Demographic Analysis: New Approaches to the Emigrant Component'. Journal of Official Statistics,Vol.19, No.4, 2003, pp. 421-448.

Redfern, P. (2004) 'An alternative view of the 2001 census and future census taking', Journal of the Royal Statistical Society: Series A,167, Part 2, pp. 209-228.
Simpson, L. (2007) 'Fixing the population: from census to population estimate'. Environment and Planning A,Vol.39, pp.1045-1057.

Simpson, L., Hobcraft, J. and King, D. (2003) 'The 2001 One Number Census and its Quality Assurance', London: Local Government Association. Available at:
www.lga.gov.uk/Documents/Publication/onenumbercensus.pdf.

Smith, J., Blackwell, L. and Lynch, K (2002) The ONS Longitudinal Study: Quality Issues from 30 years of data linkage, Journal of the UNECE, 20, pp39-49.

Statistics Canada (2003) 'Quality Management in the 2006 Canadian Census of Population'. Paper presented at Joint Statistical Meetings, San Francisco, August 2003.

Statistics Commission (2007a) Preparing for the 2001 Census – Interim Report. Report No.32, London: The Statistics Commission.

Statistics Commission (2007b) Counting on Success. The 2011 Census- Managing the Risks, Report No 36, London: The Statistics Commission.

Stuart, E. A. and Judson, D. H. (2003) 'An empirical evaluation of the use of administrative records to predict Census Day residency'. Presented at the Joint Statistical Meeting, San Francisco, Aug. 3rd–7th.

Statistics Commission (2008) A Candid Friend: Reflections on the Statistics Commission 2000-2008. Report No. 40, London: The Statistics Commission.

Statistics New Zealand (2008) 2006 Quality Management Strategy (QMS) Summary Report. Available at www.stats.govt.nz/census/about-2006-census/methodology-papers/quality-management-strategy-report.htm

Treasury Select Committee Eleventh Report of 2007-08, Counting the Population, London: The Stationery Office. Available at: www.publications.parliament.uk/pa/cm/cmtreasy.htm

UNECE (2008) 'Census Quality and Evaluation', presented to the Joint UNECE/ Eurostat meeting on Population and Housing Censuses, 13-15 May 2008, Geneva.

United Nations (2006)Principles and Recommendations for Population and Housing Censuses, Revision 2. Series M No.67/Rev2. United Nations, New York. ISBN 978-92-1-161505-0.

US Census Bureau (2003a) Evaluation of the Census 2000 Quality Assurance Philosophy and Approach Used in the Address List Development and Enumeration Operations: Final Report. Washington: U.S Census Bureau.

US Census Bureau (2003b) 'U.S. Census 2010 Quality Assurance Strategy'. Paper presented at Joint Statistical Meetings, San Francisco, August 2003.

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

Appendix A Lessons learned from 2001

1. The quality team has reviewed lessons learned from the 2001 Census and these have informed the 2011 Census Data Quality Assurance Strategy. The following list contains issues related directly to data quality assurance and was compiled from a review of ONS 2001 Census Evaluation Reports (ONS), from the Treasure Select Committee Report (TSC), Public Accounts Committee's Report (PAC), National Audit Office Report (NAO), from the Office of the Deputy Prime Minister (ODPM) and from the Local Government Association (LGA).

Table A1 Lessons learned from 2001

Source	Issue raised	Notes
ONS	The needs of customers and the feedback from them should be noted and incorporated wherever possible into the development of systems and products	2011 Data QA Strategy draws on 2001 customer issues and will explicitly seek, record and address 2011 user expectations
ODPM	There is an urgent need to improve the alignment between different sources of population data. We recommend that in the small number of authorities where there remains a problem between the council and ONS about the size of the population, following the 2001 census, a data matching exercise should be undertaken by an independent third party. This should be completed in time to feed in to next year's Local Government Finance settlement and ONS should be bound by the result. (Paragraph 44).	ONS conducted address matching exercises in 2003 which led to subsequent adjustments to mid-year population estimates. One key finding from the quality assurance studies carried out to validate the 2001 population census was the great variability in the association between administrative counts and census estimates at local authority level. One lesson was that the design of statistical sources (including the 2001 census) placed insufficient emphasis on enabling differences between sources to be understood. For this reason, analysis of administrative and other data comparators will begin with the evaluation of the 2009 Census Rehearsal and provision is made in the data QA strategy to conduct data matching exercises that will explain gaps between 2011 Census estimates and alternative sources. Consultation with LADs on the data sources they have and where they diverge with mid-year population estimates is integral to this approach. In addition, some of the new questions and changes to the definition of who should complete the form will help enable differences between sources to be understood.
LGA	The evidence and expert judgements involved in the quality	For 2011 the information used by the QA Panel in quality assuring the

	assurance of the One Number Census (ONC) should be supported with published analyses	adjusted LAD estimates will be published via the internet in LAD reports.
LGA	ONC- issues of enumeration that make population size hard to estimate should be tackled with high priority	Refinements to the coverage adjustment methodology are being made with input and advice from the UK Census Design Methodology Advisory Committee.
LGA	ONC- Key administrative sources must be assessed and calibrated not only against the census but as independent indicators of population size	Administrative sources are being assessed for triple system estimation and as comparators in QA. Endogeneity between coverage adjustment and QA will be monitored and avoided.
LGA	Development of population statistics and the census should be accompanied by a focus on the needs of the statistics users, in such a way that maintains their confidence in the methods and products.	The 2011 Census quality team is collaborating closely with the ONS Centre for Demography and developing jointly a Stakeholder Communications Strategy that aims to build confidence in the census through communications at every stage of census design, implementation and processing.
LGA	Local authority concerns with their own population estimates should be treated as opportunities to learn how population statistics can be improved greatly.	The 2009 Census Rehearsal will provide an opportunity to engage with LAD statisticians. The strategy includes consultation with all local authorities, co-ordinated through the regional framework, which will seek to learn and secure benefits from local knowledge and data.
ONS	All major problems in coded data were identified in a comprehensive check of the first EA. However, any future contract should incorporate a requirement to have a large block of data delivered early enough for checks to be carried out and corrections applied before much more of the data is coded.	QA checking will begin on the early (week 3) data batch. Prior to this, data as they pass through the suppliers' system will be sampled and checked and management information will provide early evidence of the completeness and quality of responses.
ONS	Consistency *(in data coding)* is a good measure for accuracy except where descriptions are coded consistently, but inaccurately. Checking a sample of coded data was an effective method of identifying these systematic errors, assessing reported accuracy and improving accuracy over time.	Additional checks of variable distributions and cross-tabulations of coded data against expected values will provide additional validation and identify consistently applied errors. This forms part of the early topic QA activity.

2. The quality team has conducted interviews with many ONS staff who were involved in different aspects of the 2001 census, including QA. The 2011 Census Data Quality Strategy incorporates the best practices from the 2001 exercise and seeks to avoid some of the potential pitfalls identified.

3. A review of 2001 Census correspondence from 75 local authorities (LADs) has been carried out. This review identified specific issues and concerns from each LAD. It has enabled robustness checks, to ensure that the 2011 Data Quality Assurance Strategy addresses all problems encountered in 2001.

The following issues were identified:

1. Enumeration difficulties occurred in hard to count areas where there were high numbers of multi-occupancy dwellings. The 2011 quality strategy will focus special attention on areas where enumeration problems are anticipated. See section 2.1.6 for further information on hard to count population subgroups. The management information system will provide information for coverage assessment and response rates. Qualitative information from the field will be reported back to inform the quality assurance process. Enumeration will also be improved by an up to date address register used for tracking questionnaires as outlined in section 2.2.2.3 of the strategy.

2. Discrepancies with administrative data and other non-census sources. Concerns were raised after LADs compared the census data with their own administrative records, for example council tax data and other administrative information. As stated in section 2.2 of the strategy, demographic level QA will ensure that where census population estimates and structures diverge from non-census sources, the differences are understood and can be explained, or prompt a review of census estimates. QA analysis will include comparison of population subgroups with alternative data sources.

3. Undercount. LADs expressed concern about undercount and its impact on fund allocation. LADs tended to assume that the rolled forward 2000 estimates were correct compared with the rebased 2001 estimates. Hence they assumed that the difference was explained by undercount. As with issue 2, demographic QA will ensure that differences between Census and mid-year population estimates are understood and can be explained. Other concerns were the enumeration of illegal immigrants, casual workers on student visas and overstayers and these will be addressed in 2011 in population subgroup analyses.

4. Unprocessed questionnaires. Adjustments were made to the estimates for affected LADs in 2001. For 2011 this issue will be addressed in the operational quality management plan which will ensure that procedures are in place prevent this happening again.

5. Classification issues In 1991 enumerators completed the classification of housing type on the questionnaire (though this could be

subsequently amended by the respondent). In 2001 this was completed by respondents and there were 1991/2001 inconsistencies. There will be 2001/2011 continuity in that respondents will again complete this classification. Further inconsistencies arose through automated data processing for example where an industrial area (household dwellings with no usual residents) was coded to second/ holiday homes. Comparator datasets will identify instances of misclassification, as outlined in section 2.1.2 of the strategy.

6. Output table inconsistencies. Some LADs identified inconsistencies in output tables and this was mostly due to small cell adjustment. Improbable variable combinations were also identified, for example young people listed as retired. Quality assurance analysis will include checks for implausible characteristics in multivariate tables/ analysis.

7. Mobile population issues. LADs with special populations such as students and armed forces personnel expressed concern about under-enumeration of these particular groups. Again, the 2011 quality strategy will focus special attention on areas where enumeration problems are anticipated. Data will be obtained from the Higher Education Statistics Agency (HESA) and the Defence Analytical Services Agency (DASA) to validate numbers of students and armed forces personnel respectively (see Appendices E, G and H for data sources used in 2001 and datasets considered for 2011 quality assurance).

8. Estimation issues. Following 2001, the ONS matching studies highlighted limitations in the Hard to Count (HtC) index and strata for the census coverage survey (CCS) sampling. These issues are being addressed in revisions to the design of the coverage survey and adjustment.

9. Low response rate. This relates to a number of factors including enumerator difficulties and refusals. Again, response rates will be managed via the management information system and special attention given to areas where enumeration problems are anticipated.

10. Timeliness of census data release. The Data QA Strategy prioritises variables that support key uses of census demographic information, in particular mid-year population estimation as outlined in section 2.1.1.

Appendix B Data quality monitoring system: interfaces with other data producers and systems

Table B1

System	DQMS requirement
Management information system	To receive and interrogate management information on the progress of field and data capture operations prior to census data being available for analysis and to inform the QA Panel of issues emerging from field operations.
Image management system	To examine census form images to aid understanding of data error.
Audit trail information	To identify the points in census processing at which records, either individually or as aggregates, have been changed; by which process and the nature of the change.
Address register	Source against which questionnaire tracking information (in the MI), census data and administrative sources can be checked. To use evidence of address register additions and amendments as indicators of data quality to inform QA Panel decisions.
ONS geography	To assign different geographies to the data and comparators.
Capture and coding team.	To consider errors identified through consistency checking and to ensure that QA activities are synchronised and meet supplier acceptance/ rejection decision deadlines
Edit & imputation team	Joint users of QA reports
Census coverage team	To receive coverage estimates and post-imputation data. To access data used in coverage estimation.
Population estimation/ ONS Centre for Demography	To receive population estimates and measures of population based on administrative sources. To receive migration estimates from surveys and administrative sources. To incorporate demographic analysis (fertility, mortality and sex ratios) in QA and to support ONS Centre for Demography involvement in the QA process.
Neighbourhood Statistics (NeSS)	To receive NeSS data as comparators. To support QA usage of NeSS data quality management methodology and tools. To support NeSS input to the QA process.
ONS Longitudinal study	To receive ONS inputs to the QA process on underenumeration (modelling of longitudinal data), on overenumeration (evidence and data on multiple enumerations) and to receive information on quality issues emerging through record matching.
Administrative Sources Integration Directorate	To receive newly acquired administrative sources as comparator data and to access details of their quality analyses of those sources. To input to the QA of those data sources.
Census outputs system	QA packs for each LAD for web publication, census quality reports for web publication and for Eurostat, Statistics Board Assessment Centre. To facilitate efficiency and consistency with other census outputs.
Census metadata system	To report quality of census data. To access provisional census metadata systems to support QA.
Data visualisation team	To access the software, expertise and experience of the data visualisation team to highlight 'hotspots' through visual assistance.

Appendix C Prioritised data quality assurance tasks

Priority*	Task
M	Create prioritised list of LADs (2.1.2)
M	Identify, analyse and prepare rehearsal comparators (2.1.2)
M	Identify, analyse and prepare census comparators (2.1.2)
M	Develop data visualisation templates for automated, R(ed) A(mber) G(reen) analysis (rehearsal) (2.1.2)
M	Develop data visualisation templates for automated, RAG analysis (census) (2.1.2)
M	Develop and confirm topic QA checking framework (2.1.2)
M	Identify topic QA checks for rehearsal (2.1.2)
S	Develop cumulative checks- topic QA (2.1.4)
S	Develop cumulative checks- demographic QA (2.1.4)
C	Make rehearsal data available to topic experts for QA (2.1.5)
C	Make census data available to topic experts for QA (2.1.5
S	Secure topic expert input to identify comparators, set tolerances and QA data (2.1.5)
S	Identify, compare and monitor population subgroup counts and characteristics at all geographic levels (2.1.6)
S	Identify and monitor problematic areas for investigation (2.1.6)
S	Repeat topic checks on IDC data to understand and monitor modal bias (2.1.7)
M	Create expected population distributions for each LAD (2.2)
M	Identify, develop and implement demographic indicators for each LAD (2.2)
S	Demographic QA: develop cumulative checks of characteristics and demographic indicators (2.2)
S	Review demographic comparators ahead of census (2.2)
S	Identify and monitor dummy form volumes by type at all geographic levels (2.2.1)
S	Single year of age and post coverage adjustment post imputation checks (2.2.2)
S	QA overcount adjustments at all geographic levels (2.2.2)
W	Cohort analysis using 2001 including student ratios and age-related covariate analysis (2.2.2.2)
S	Monitor volumes and patterns of addresses with no response and no dummy form (2.2.2.3)
C	Use administrative sources to validate data for larger CEs (2.2.2.3)
C	Cross-validate coverage adjustment and address register (2.2.2.3)
C	2001/2011 LSOA comparison and adjustment (2.2.2.3)
S	Produce and provide cumulative, regional and national distributions and demographic indicator data for QA Panel (2.2.3)
S	Secure local data and evidence from LADs via regional statisticians (2.2.3)
S	Reconcile LS evidence and coverage adjustments including multiple enumeration adjustments (2.2.4)
S	QA overcount and population base adjustments from a variety of sources (2.2.4)

M	Collect and analyse evidence on England and Wales-born people living abroad (2.2.4)
M	Produce LAD packs for QA Panel (2.2.5)
S	Produce LAD packs for publication (2.2.5)
S	Analyse and collate additional data and evidence for 'problematic' LADs (2.2.5)
S	Analyse and include LSOA data for LAD packs (2.2.5)
S	Agree comparator data quality checklist with stakeholders (2.2.6.1)
S	Test and evaluate comparators drawn from new administrative sources (2.2.6.1)
C	2001 LS/CCS linkage (2.2.6.2)
C	2011 LS/CCS linkage (2.2.6.2)
S	Administrative data linkage to resolve quality queries in use of administrative data (2.2.6.2)
S	Analyse cohort sex ratio trends over time (2.2.6.3)
M	Collect and collate supporting information (eg from field and LADs) for the QA Panel (2.2.7)
S	Maintain and monitor LAD and cumulative totals of people imputed by coverage adjustment (2.2.9)
M	Stakeholder engagement in line with User Communications Strategy (2.2.10)
M	Ad-hoc analysis requested by QA Panel (2.2.11)
M	Supplementary analysis: planning and implementation (2.2.12)
M	Appoint and support QA Panel (3.1)
S	Maintain QA Working Group (3.2.1)
S	Establish and support UKCDMAC QA subgroup (3.2.2)
C	Submit strategy/ plans for international peer review (3.2.3)
M	Produce data quality report for ONS publication (3.4)
M	Produce data quality report for Eurostat (3.4)
S	Prepare material for Statistics Board Assessment (3.4)
S	Conduct rehearsal QA (4)
S	Identify and quantify population subgroups for population base adjustments
C	Investigate and implement spatial analysis for QA
C	Set up and maintain spatial analysis subgroup

*M=must, S= should, C= could, W= won't

Appendix D Indicative checking framework for topic-QA

Table D1 Checking framework for topic QA of first LADs processed

Process prior to check:	Data load	Reconcile multiple responses within HH*	Apply filter rules	Item imputation	Post coverage imputation	Assign output geographies	Disclosure control adjustment
Checks to be performed, by variables checked							
Compare with expected values	All applicable variables	All if more than 5 per cent (tbc) within LADs are duplicates	No	All applicable variables including migrant and 2nd residence data Migrants Second Address	Age*sex, Sex* ethnic group, age* economic activity, sex* marital status. Include migrants and 2nd residence checks. Age*sex*TTIND Relationship Migrants Second Address	For OAs, compare with 2001 age*sex, sex* ethnic group and age* economic activity, sex* marital status	All applicable variables and key complex derived variables
Statistical distance from comparator	All applicable variables	All applicable variables	All applicable variables	All applicable variables	All applicable variables	N/a	All applicable
Cross-tabulation	Age * sex Age* marital status Age* student indicator Age * Economic activity Sex* occupation	No	No	All affected variables – check for implausible combinations	Age * sex Age* marital status Age* student indicator Age * Economic activity Sex* occupation	No	All affected variables – check for implausible combinations
Ad hoc		Compare levels and characteristics of duplicates with 2001 Census, 2009 Rehearsal and LS Check household structures.	Compare 2001/ 2011 missingness for each variable. Analyse within-record missingness.		Comparison of imputed population with coverage estimation estimates. Compare pre- and post- adjustment household structures with 2001 patterns.	Identify characteristics of change and refer on to Demographic QA.	Compare pre- and post-adjustment household structures.

* This process identifies where there is more than one form for an individual within a household.

Later data batches will receive fewer checks than the earlier batches. An indicative checking framework for the final LADs processed, summarised in Table C2, suggests that only 1 in 10 LADs have the same level of validation as the early LADs and the scale of other checks will be reduced. For planning purposes it is estimated that around 12 LADs will be fully checked as set out in Table E1 and the remaining LADs will be checked according to the schema in Table E2.

Table D2 Checking framework for topic QA of final LADs processed

Process prior to check:	Data load	Reconcile multiple responses within HH	Apply filter rules	Item imputation	Post coverage imputation	Assign output geographies	Disclosure control adjustment
Checks to be performed, by variables checked							
Compare with expected values	For 1/10 LADs: All applicable variables. Remaining LADs: Sex, Age, Ethnic Group, Student status, term-time indicator, marital status.	All if more than 5 per cent (tbc) within LADs are duplicates	No	Imputation rates for all LADs but examination of only problematic ones Migrants Second Address	Age*sex, Sex* ethnic group, age* economic activity, sex* marital status. Include migrants and 2nd residence checks. Age*sex* TTIND Relation-ship Migrants Second Address	For OAs, compare with 2001 age*sex, sex* ethnic group and age* economic activity, sex* marital status	For 1/10 LADs: All Applicable variables and key complex DVs. Remaining LADs: Sex, Age, Ethnic Group, Student status, term-time indicator, marital status.
Statistical distance from comparator	All applicable variables	No	No	No	No	N/a	All applicable variables
Cross-tabulation	1/10 LADs: Age * sex Age* marital status Age* student indicator Age * Economic activity Sex* occupation	No	No	No	No	No	1/10 LADs: Age * sex Age* marital status Age* student indicator Age * Economic activity Sex* occupation
Ad hoc		For 1/10 LADs: Check household structures..	For 1/10 LADs: Compare 2001/ 2011 missingness for each variable.			Identify characteristics of change and refer on to Demographic QA.	Compare pre- and post-adjustment household structures.

Note: Where a sample of LADs is selected for more extensive checking, these will be LADs classified by demographic QA processing as 'hard-to-count'

These respective proportions are subject to review, pending checking outcomes. This staged approach has the advantage of freeing up QA analysts who will be familiar with 2011 Census data for diversion onto demographic QA validation.

The second tranche of census data received in week 42 will include information returned on paper forms after the tenth week following census day. It is estimated that there will be around five per cent of people enumerated in each LAD in this final batch of data. Where that assumption is challenged, or where the addition of the second tranche of data creates significant deviations from existing LAD distributions, this will be further analysed by the quality team.

ONS must accept or reject data supplied by the contractor within 12 weeks of receipt. Most data will be supplied in batches of local authority districts and most checks will be LAD-based. The availability of data for checking purposes is summarised in Table E3.

Throughout census field operations, management information will be available for ONS to monitor and manage the various census processes. This will include questionnaire tracking information which provides proxy data on response rates and household counts and will be monitored by the quality team ahead of the main stages of data quality assurance.

During the first three weeks after census day, extracts of data from the suppliers' data systems can be used to run preliminary checks.

A preliminary batch of data received in week 3 will have passed through all capture and coding processes though it will be incomplete because it will only include census returns received and processed up to week 3. This limits the scope for checking against comparator sources. Tolerances will be set to identify spikes that indicate systematic error. This is perhaps the most important stage of topic QA since there is scope to adjust census process to correct data error early on in the processing cycle.

From week 12, data will be supplied to ONS in groups of LADs, the grouping and order of delivery to be specified by ONS. The processing order will be informed by the imperatives of data matching and specifically the need for contiguous LADs for census/coverage survey matching. These data will have passed through all capture and coding processes and will be checked for customer acceptance. In parallel with the data QA processes described here, the capture and coding team, separately from the QA team, will be completing consistency checks on these data, with reference to coding schedules and census form images.

Table D3 Data availability for QA purposes					
Date*	Before W-3	W-3 to W3	W3	W12-W39	W42
	MI reports	Supplier extracts	Preliminary batch	LAD batches	Wk 10+ data
Data available	Management information reports	Extracts from Supplier processing. Possibly associated form images.	Captured and coded data and associated form images	Captured and coded data and associated form images	Captured and coded data and associated form images
Data coverage (relative to census day)	Questionnaire tracking information and field reports	Data supplied on forms prior to the point of extraction	Data supplied on forms returned by week 2	Data supplied on forms returned up to 10 weeks after census day	Data supplied on forms returned following 10 weeks after census day
Geographic coverage	All LADs	All or selected LADs	1 batch of 3 LADs	In pre-specified batches of LADs	All LADs

*Dates are given with reference to census day, so W-3 is three weeks before census day; W10 is 10 weeks after census day.

Appendix E Sources of data for population subgroup analysis used in 2001

In 2001, the following datasets were used for comparison of subgroups:-

Table E1 Sources of 2001 comparator datasets for subgroup analysis

Subgroup	Data source
0-year olds	Birth registrations data
Over-90s	Information on claims of state retirement pensions and/or other benefits from the Department for Work and Pensions (DWP)
Students	Information from the Higher Education Statistics Agency (HESA) and the Learning & Skills Council (LSC)?
Home armed forces	Counts provided by the Defence Analytical Services Agency (DASA)
Foreign armed forces and their dependents	Information provided by the United States Air force (USAF)
Prisoners	Information provided by the Home Office

A list of datasets being considered for 2011 quality assurance activities, including population subgroup analysis, is shown in Appendix H

Appendix F Multiple enumerations in 1991

Analysis of 1991 visitor information from the ONS longitudinal study in Table A3 illustrates the likely combinations of residence/ absence/ visitor status and their respective attribution as primary or secondary records. The LS is a representative sample of 1 per cent of census records. This is necessary so that their inclusion or omission from population bases can be made. Multiple enumeration is likely to be at higher levels than in 2011 as a result of more complex lifestyle patterns.

Table F Multiple enumeration among present residents, absent residents and visitors in 1991	
Primary record is a present resident record	
Enumerated once as present residents	95.40
Enumerated twice as present residents	0.07
Enumerated twice as present resident and absent resident	0.13
Enumerated twice as present resident and visitor	0.16
Primary record is a visitor record	
Enumerated once as a visitor with no matching resident record	1.84
Enumerated twice as visitors	0.06
Enumerated twice as visitor and present resident	0.17
Enumerated twice as visitor and absent resident	1.29
Primary record is an absent resident record	
Enumerated once as absent resident	0.86
Enumerated twice as absent resident	0.01
Enumerated twice as absent resident and present resident	0.01
Enumerated twice as absent resident and visitor	0.00
Total sample	543834

Source: ONS longitudinal study

Appendix G Datasets used for 2001 quality assurance

In 2001, the following datasets were used as comparators:

- published rolled-forward Mid-year Estimates (ONS)
- 2000 Mid-year Estimates extrapolated to mid-2001 based upon average annual changes (ONS)
- estimates of under ones from birth registration data adjusted for infant mortality and migration (ONS)
- patient registers recording the number of patients registered with NHS GPs (ONS via health authorities)
- people aged 65+ claiming state retirement pension and/or other benefits (DWP)
- children aged under 16 receiving child benefit (DWP)
- children aged 5-14 attending educational establishments (DCSF and WA)

Datasets used in 2001 for population subgroup analysis are shown in Appendix E

Appendix H Additional datasets to be considered for 2011 quality assurance

In addition to the datasets shown in Appendices E and G, the following data sources are being considered for 2011:

- Valuation Office Agency - Council Tax data including Council Tax list and non-domestic rating.
- Department for Constitutional Affairs - CORE (co-ordinated online record of electors).
- Department for Communities and Local Government - National Register of Social Housing.
- Department for Children, Schools and Families School Census.
- Wales School Census (Welsh Assembly Government)
- Independent Schools Council – Numbers on roll in independent schools
- Children's Act 2004 information database.
- Learning and Skills Council - information on students.
- Home Office - National Identity Register, data on asylum seekers and refugees.
- Commission for Social Care and Inspection - CSCI database -details of care establishments.
- Department for Work and Pensions - Work and Pensions Longitudinal Study, the Migrant Worker Scan and the National Insurance Number allocations data
- Department of Health/Connecting for Health - National Health Service Central Register, NHAIS and successor systems - Population Demographic System (PDS), SUS and back office information.
- HM Revenue and Customs - data on income (PAYE and self- assessment) and information on Child Benefit.
- Laing and Buisson Care Homes Database
- National Insurance records
- Passport data
- International Passenger Survey

Additional datasets for investigating population subgroups will be identified during 2009.

Appendix I Data QA checks in the 2009 Census Rehearsal

The Census Rehearsal will be an opportunity to:

- test and confirm the utility and effectiveness of pre-programmed data validation checks
- practice and develop ad-hoc analytic skills to supplement automated checks
- test the quality of comparator datasets against rehearsal and address register data and against administrative sources provided by the rehearsal local authorities (and explore triangulating between these three sets of data)
- test and develop the data quality monitoring system and the associated software used in QA.
- test and evaluate the potential for early QA analyses to inform field activities e.g. targeted publicity to low responding population subgroups
- review and evaluate procedures for the co-ordinated acceptance of data batches following quality assurance and data validation/consistency checks
- enable the quality team to develop effective working relationships with key stakeholders including LADs and topic experts

The following checks will be carried out on all three rehearsal LADs.

Extracts of census data will be drawn from the supplier's system prior to data coding to gain an early assessment of the quality of key demographic data fields. An example of the data fields examined and their checks are given in Table I1.

Table I1: QA checks on early extract data

Question	Check
Household q1 (Usual resident categories H1)	Frequency of single tick and multi-tick responses by type compared against expected values
	Cross-validation between tick 3 (students away from home) and q6 in the individual questions for this form
	Cross-validation between final tick (second address/holiday home), H4 responses and q12 on the individual questions associated with this form.
Individual qs 12 and 13	Cross-validate these two responses for each form
	Sex ratios compared with expectations.
	Validation checks on priority variables:- Marital Status Ethnic Group Students Term time indicator

Topic QA will be restricted to the following variables: sex, date of birth, marital/ civil partnership status, ethnic group, student status, term-time indicator, number of people in the household and tenure. The QA team will consult the following topic leads in the rehearsal QA:

- ethnicity/identity/language/religion
- housing
- migration, population bases and definitions
- demographics and social composition

Data for these variables will be compared against expected values after they have been loaded into the downstream processing system.

Within each LAD, the following multivariate analyses will check for implausible variable combinations:

Age x sex x marital/ civil partnership status x ethnic group
Age x sex x student status x no. people in HH x tenure x term-time indicator

Data for the following population subgroups will be compared against expected values (in all areas unless indicated otherwise):

- residents in communal establishments
- armed forces personnel
- prisoners
- students
- migrants, both long and short-term
- people living on caravan sites (Anglesey and Lancaster)
- people living on travellers' sites (Newham)

This analysis will use data sources newly available to ONS for QA purposes, including the DWP Migrant Worker Scan data and HESA information.

Sex ratios by age and geography will be analysed for each LAD and as a cumulative total. The topic QA and sex ratio analyses will be supported by data visualisation techniques.

The rehearsal will provide an opportunity to QA the address register in areas where either expected or observed response rates are low, by comparing the Register against alternative sources available from local authorities, including counts of properties and sub-properties.

Responses by age and sex for each LAD will be compared against an expected range of values drawn from a range of sources, including:

Appendix J Timeline for data quality assurance activities

	2008				2009				2010				2011				2012				2013
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1
Create and Publish Version 1of QA Strategy (Apr 08- Mar 09)																					
Topic Lead Collaboration (Feb 09-Mar 2013)																					
Identify and prepare comparators (Oct 08-Sep 09/Sep 09-Feb 11)																					
Rehearsal and evaluation (Jul 09- Mar 2010)																					
Census Topic QA (March 2011- May 2012)																					
Census Demographic QA (July 2011-Aug 2012)																					
QA Panel Involvement (July 2010-May 2012) Contingency (April-- July 2012)																					
Supplementary analysis (Apr 2010-Mar 2013)																					
Preparation and checking of Census estimates for Publication (Aug 2012-Mar 2013)																					