Distr.
GENERAL

CES/SEM.40/9
14 September 1998

Original: ENGLISH

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

CONFERENCE OF EUROPEAN STATISTICIANS

Joint ECE-EUROSTAT work session
on Population and Housing Censuses[1]
(Dublin, Ireland, 9-11 November 1998)

Study topic 2

QUALITY CONTROL AND RESULTS VALIDATION

Invited paper submitted by the Office for National Statistics, United Kingdom[2]

I. Introduction

1. There are many sources of potential error in the processing of census data. This paper focuses on the role that quality control and results validation will play in managing these sources of potential error for the 2001 Census of Population for the United Kingdom.

2. Quality control and quality management are fundamentally different approaches to managing the quality of a process. Quality control is implemented to ensure that a minimum acceptable quality standard is met for every unit of output from a process. Quality management assesses the quality of a sample of the output of a process, and aims to continue to improve the average quality of output from that process over time through implementing improvements to the process. Census processing is best served by a combination of the two approaches.

3. Quality management does not generally involve reworking to correct errors, whereas quality control ensures that certain errors do not exist for any unit of output. While quality management is an important part of our

[1] The papers which are prepared for this work session will be treated in the same manner as papers that are prepared for seminars.

[2] Prepared by Nick Parsons, Data Quality, Census Division.

management strategy for the 2001 Census, the main focus of this paper is quality control.

4. Validation of results will aim to identify systematic errors in thedata which have not been identified through quality control or quality management.

5. Our objective is to minimise the level of significant systematic error in the 2001 Census data through a combination of quality control, quality management and the validation of results.

6. The main processes are outlined in Section 2, to provide some context for the rest of the paper. Section 3 considers the need for quality control of processing and some of the quality control measures we plan to implement for 2001. Section 4 provides a brief description of the approach we will take to quality management in the clerical processes and where we will adopt a quality management approach in the automated processes. The role that validation of results will play in assuring the quality of 2001 Census data is described in Section 5.

II. Background: Outline of Processing

7. The 2001 Census will be conducted by a delivery - post back method. An Enumerator is responsible for a unit of work called an Enumeration District (ED) consisting of about 200 addresses. They will identify and deliver forms to all households within their ED. Each household withinan ED will be uniquely identified by a Record Number (RNO).

8. After Census night, Enumerators will follow-up where a form has not been posted back for a household.

9. Forms posted back will not be physically merged with those received from the field operation through the follow-up of forms not posted back. The unit of work throughout most of processing will be the form (or batches of forms) rather than the ED. Data for an ED will be merged before the Edit and Imputation process.

Receipt

10. This process will register the forms received by the processing centre. Forms will arrive at the processing centre from two sources; those posted back by the public and those collected by the Enumerators through follow-up of forms not posted back.

Scanning

11. Scanning and Recognition are the two processes we will use to capture the responses from the census forms. Scanning is the process in which an image of each page of a form is created. These images are then used for both automatic and clerical interface processes, replacing the physical forms.

Recognition

12. Recognition is the process in which ticks and write-in responses are captured from the images created in Scanning, and used to create the first cut of electronic census data.

13. The questions on the 2001 Census form are of three types; tick-box questions, combination tick-box / write-in questions, and write-in only questions. Ticks by respondents to the tick-box and combination questions are captured through Optical Mark Recognition (OMR). Write-in responses, including those to the combination questions, are captured through Optical Character Recognition (OCR).

Automatic Coding

14. Write-in responses captured through OCR as part of the Recognition process are compared against a list of responses. Where a match is achieved, the code associated with that response on the list is assigned to the response from Recognition.

Query Resolution and Expert Coding

15. Responses which are not coded through Automatic Coding, will be coded in Query Resolution. A person will code the response via an interface that incorporates an image of the response and a computer-assisted coding package. Responses that cannot be coded in Query Resolution will be coded in Expert Coding using different procedures or reference material.

Edit and Imputation

16. Editing will be largely automated. The Edit process will resolve cases where a respondent has marked multiple tick-boxes for the one question, and will identify, for Imputation, data items which fail consistency checks.

17. Imputation will assign values for all items where a response to the related census question was not given. Imputation will also resolvecases which fail the consistency checks, by substituting non-conflicting values for one or more of the items in conflict.

Post Edit and Imputation Processes

18. The feasibility of adjusting the 2001 Census database, using information on coverage obtained through a post-census Coverage Survey, is being considered. This is known as the One Number Census project.

19. Disclosure Control, designed to prevent individuals being identified through the census data, will be implemented before the 2001 Census data is released to Census Output for dissemination.

Clerical Processes

20. People have different skills and experiences, and will interpret or perceive the same response differently. This is particularly the case with write-in responses to census questions which involve concepts that are complicated or unfamiliar to the person coding the response. There will also be a level of random error introduced by people working in clerical processes.

21. These two factors mean that there is a level of variance introduced by the different people working in a clerical process. We can manage this variance through quality management, but ultimately there will always be some variance associated with a clerical process.

22. Systematic errors or bias can also be introduced through the systems and procedures used by people working in a clerical process. It is these systematic errors which are most significant to our customers. One advantage of a clerical process is that staff working in the process can potentially identify these systematic errors in the course of their work.

Automated Processes

23. The processing of the 2001 Census of Population for the United Kingdom will be significantly more automated than the processing methodology used in 1991. The introduction of OMR technology will allow approximately 90% of "simple" (for example Country of Birth) coding to be performed automatically. Approximately 60% of “complex” (for example Occupation) coding will also be completed automatically using OCR and Automatic Coding technology.

24. There are obvious and substantial cost advantages in increasing the level of automation in a processing methodology. There are also some implications for managing the quality of the data.

25. Automation greatly increases the consistency of processing. The same response will always be coded the same way in an automatic process. While this consistency is positive if the automatic process is coding the response correctly, it can also produce systematic errors if the automatic process is coding the response incorrectly. These systematic errors can be very obvious, and far more significant to users of census data than the variance associated with a clerical process.

26. We will use a combination of quality management, quality control, and results validation to minimise the risk of systematic errors in the data we release to our customers.

III. Quality Control

27. One of the major requirements of census data is that it provides accurate data at the small area level. In the United Kingdom, census data at the Local Authority Area level is used by the central government as the basis for

distributing resources, and for defining electoral boundaries. Many other customers use census data at smaller area levels.

28. Quality control is important for census data because what we produce is a differentiated product, meaning that every unit of output is different. If this were not the case, we could rely entirely on quality management, and simply replace any defective units that we produce. Processing census data is in many ways like a production line, but in this fundamental sense it is a quite different operation.

29. The census data for each unit of output must stand alone, achieving at least a minimum, acceptable standard of quality. We will implement quality control to ensure the data we produce meets this standard.

Scanning and Recognition

30. One of the key risk areas for any census is the capture of the data, the transition from paper forms to the first cut of electronic census data. We will apply quality control extensively to ensure the quality of the imaging of the forms by Scanning, and the capture of responses from these images by Recognition.

31. There is essentially no difference to our customers whether a household or person fails to be enumerated by the field operation, or is enumerated by the field operation but does not make it into the census data during processing. The end result to the customer is the same in both cases, those households or people are missing from the census data.

32. Once the forms are dispatched from the field to the processing centre, errors can be introduced through a range of sources. These include errors in the Scanning process (for example forms not scanned) and errors in the Recognition process (for example ED identifiers for a form recognised incorrectly).

33. Ensuring the integrity of the data captured from the census forms is essential. We will implement quality control to achieve this.

Data for Every ED

34. Checking that we have forms for every ED will be an important task in the management of the field operation.

35. We also need to ensure that we have data for each and every ED within the census geographic hierarchy that defines the EDs for which Census Output expect data for dissemination.

Data for Every Household

36. Enumerators will be issued with an Enumerator Record Book (ERB), which they will use to record information about the households they deliver census forms to. This information will include the Record Number (RNO) they assign to each household to uniquely identify it within the ED.

37. Some forms will be posted back directly by the public. Those that are not posted back will be collected by the Enumerator and forwarded to the processing centre.

38. Some form of “Field Reconciliation Check", between the RNOs for which we expect forms, and the RNOs in the data created for that ED, is required. This Field Reconciliation Check will be implemented at the completion of Scanning and Recognition. The RNOs expected (from the ERB) will be used as the control information.

39. Procedures will be developed to resolve cases where RNOs expected for an ED are not in the data, where RNOs not expected are in the data, and where more than one form has been captured for a RNO. These procedures will allow for adjustment of the control information, and for additions, deletions or modifications to the captured data.

Data for Every Person

40. There will be a significant proportion of households which do not come into contact with an Enumerator at any stage. A common case will be where nobody is at home when the Enumerator calls during delivery, and the form for that household is posted back promptly. For these households there will be no information from the Enumerator about how many persons to expect for that RNO.

41. This means that the Enumerator can not provide control information in the way that they can about RNOs expected. Another form of quality control is required if we are to ensure that every person who has provided responses on each household form is included in the data.

42. We are currently investigating the potential of using other information contained on the form, in a check against the number of persons in the data for a household.

43. The 2001 Census form includes a table which asks a member of the household to list the names of all the usual residents of that household. The count of names in this table is one of the options we are considering as a control figure.

44. Procedures will be developed to resolve cases where the count of persons in the data is different from the control figure. These procedures will allow for the insertion or deletion of person records for the household.

45. If we decide on a clerical process, this could be readily included in the Field Reconciliation Check described above. The potential benefits of a clerical process include using the name to make an intuitive decision about the sex of a person missing from the data.

46. Alternatively, if we decide on an automated, rule-based solution, persons could be automatically added or deleted from the household.

Complete Images and Data for Every Form

47. The Scanning process needs to ensure that all the required pages are imaged for each form that is scanned.

48. Procedures will be developed which force the rescanning of forms for which all pages are not imaged.

49. After the Recognition process we will check that every record created in the data is complete. Procedures will be developed which allow us to re-recognise forms where although all the required images for a form have been created, complete data for that form has not been produced by the Recognition process.

The Right Records for Each Form

50. The discussion above refers to households and person records. This is a simplification. In fact we will have a number of other different types of records, including a record for “communal establishments” (these include hotels and hospitals).

51. There are also other records which contain information about the ED and information about families within households.

52. There are conventions that govern which records are required for a household or communal establishment. The details are not important to this discussion. The point is that we need to ensure we have the correct records for each RNO following data capture.

53. We will do this by checking the data following data capture to identify cases where these conventions are not satisfied. We will also develop procedures that allow for the addition, deletion and modification of records in the captured data to ensure that these conventions are satisfied.

54. The resolution of RNOs failing this check could be readily included in the Field Reconciliation Check.

After Data Capture

55. We will ensure the integrity of the structure of the data following Scanning and Recognition through the checks and resolution processes described above.

56. We then need to ensure that integrity is not compromised in subsequent processes.

57. We will store information about the data structure for the ED following these checks. This information will then be used as control data against which we can check the structure of the data after later processes.

Editing

58. The Edit process can play a significant role in quality control of census data by applying rules which ensure the internal consistency of the data. We have considered the requirements of our customers in determining how we should define internal consistency.

59. In defining these edits, we are enforcing some conditions and conversely preventing others. Some of these rules check the consistency of data items within the one record. Other rules check the consistency of data items within a number of related records.

60. An example of an inter-record edit is one which ensures that a person under 16 years of age does not have data for the labour force related data items such as Occupation. We apply this edit because our definition of a person who is economically active includes the condition that they are over 15 years of age. We enforce the condition through the edit.

61. An example of an intra-record edit is one that checks the consistency of age, relationship, marital status and sex responses for persons within the one household. We will apply rules that enforce the condition that there is a minimum generation gap of 13 years between a parent and their children. While there may be real cases of people having children at younger than 13 years of age within the population, we have decided to apply this rule in order to be consistent with previous censuses and other surveys.

62. The Edit rules will also resolve cases of multiple responses to the one census question. We will consider common multiple response patterns in the design of these edits. Our aim is to minimise the impact of respondent error to questions, where this error involves multiple response to a tick-box or combination tick-box / write-in question.

63. We will use a combination of “hard” and “soft” edits. A hard edit is one where we force a change in the data so that the rule is satisfied. A soft edit is one where we store a count of the number of times this edit rule is not satisfied within the data, but do not force a change.

64. The counts of edit failures are a valuable source of information for quality management. An unexpectedly high or low count for a particular edit may indicate either a problem with responses or one of our earlier processes.

Range Checks

65. For the purposes of dissemination of census data, there is a clearly defined set of values expected by Census Output for each data item. This set of values generally, but not always equates to the output classification for that data item. The purpose of the Range Check is to identify any cases where the value for a data item is not in the set of “legal” values for that data item. These cases are errors in the data, and will be corrected before the census data is released to Census Output for dissemination.

66. For example, the set of “legal” values for Sex would be Male and Female, or more precisely, the numbers used in the data set to denote male and female. Let’s say that males are stored as “1” in the data, and females are stored as “2”. For the Sex variable our Range Check would identify any case where a person had a value that was not “1” or “2”.

67. The Range Check will be applied to the data after the completion of the final process that will change the data before release to Census Output.

68. A modified Range Check could also be applied at the completion of earlier processes. Different sets of “legal” values will be required where a data item can legitimately have a value in an earlier process that would not be considered "legal” by Census Output. The advantage of applying a Range Check earlier than at the completion of processing a subset of the data is that we will identify errors being introduced earlier, affording us more time to evaluate and address these errors. As reprocessing can be very costly if required on a large scale, this early detection of errors may end up being very cost effective.

Transferring Data to Output

69. The other key risk area we need to address through quality control is the transfer of the data to Output. This transfer will involve the processing centre producing a file or series of files which Census Output will then load into the tabulation system they will use for dissemination.

70. There are likely to be substantial differences between the way data is stored for processing and the way that a tabulation package will require the data to be stored for efficient tabulation. Reformatting the data to the format required by the tabulation package can potentially introduce significant systematic errors into the data.

71. We will develop a quality control check which will compare the files produced for Output with the processing database from which the file was produced. This check will identify any inconsistencies between the two sets of data.

72. Census Output will then be responsible for ensuring the integrity of the downloading of the files from the processing centre into the tabulation package.

IV. Quality Management

73. We will complement quality control with quality management to address other elements of quality. This section describes some of the key quality issues we will address through quality management, and outlines our approach to quality management of clerical processes.

The Link between Images and Processing Applications

74. The link between images and our applications is critical to the integrity of the data because we will use images as a substitute for the physical forms after Scanning. The output of the Scanning process is an image of each page of the form. It may be that images of specific response areas on the form are produced rather than images of the whole page, but the principles of the following will still apply.

75. The risk is that the census data will be created or updated by referencing the wrong images. If this were to happen systematically throughout a process, the census data could be fundamentally corrupted as a result.

76. The link between images and data does not need to be tested continuously throughout processing. It will be tested thoroughly during development of our processing applications that access images, whether for an automatic or clerical interface process.

77. To check that our processing applications are accessing the correct images, we will scan test forms, and check that the images used by the application correspond to the appropriate response on the relevant test form.

The Link between Processing Applications and Data

78. It is also critical we ensure processing applications are updating the census data correctly. As with the link between images and our processing applications, if the census data is systematically updated incorrectly by an application, the data could be fundamentally corrupted.

79. This link will be thoroughly tested throughout development of applications to ensure that the correct data item within the census data is being updated.

Recognition of Ticks (OMR)

80. The Optical Mark Recognition (OMR) of tick responses to the tick-box and combination tick-box / write-in questions will be used to code most of the data for the 2001 Census. Recognition of “phantom”ticks, or the non-

recognition of tick responses can introduce significant, systematic errors in the data.

81. We can not prevent errors in the recognition of tick responses entirely, but we can introduce quality management that will reduce the risk of systematic errors.

82. The scanners and computers we will use to create images and recognise ticks, like any electrical equipment, are subject to error.

83. Regular preventative maintenance will be implemented for the scanners to optimise the quality of the images created. The quality of each image will be checked automatically to ensure the image is suitable for Recognition.

84. We will also scan and recognise test forms on a regular basis to identify problems in the Scanning and Recognition of tick responses.

Recognition of Text (OCR)

85. The output of Optical Character Recognition (OCR) is a recognised string of text. These text strings are the input into the Automatic Coding process. Some “character repair” of characters that could not be recognised through OCR may be employed to allow more complete text strings to be passed on to Automatic Coding. This extra process will not change the relevance of what follows.

86. The key risk associated with the OCR process is the introduction of systematic “substitution” errors. These are where one character is systematically recognised incorrectly as another character. There will always be some substitution errors in the OCR of hand written responses. As with the OMR of tick responses, this is a source of error we will manage.

87. OCR software allows us to set confidence levels which the recognition process must meet before a character is considered to be recognised. Increasing these confidence levels will allow fewer substitution errors in the recognised text strings. There is an obvious trade off here between efficiency (recognising as many characters as possible in OCR) and quality (the risk of introducing errors through recognising characters incorrectly).

88. Errors in the recognition of text responses can be identified in a number of ways. One of these is to check the output of the OCR process directly by comparing the image of a response and the text string which OCR has produced for that response. Output from this check will allow us to identify where letters or numbers are being systematically recognised incorrectly by the OCR software.

89. Another way to identify substitution errors is to include this indirectly as part of quality management of Automatic Coding. We will check the output of Automatic Coding by reprocessing a sample of automatically coded responses through the Query Resolution process, and comparing the results. We could include a check on the recognised text string from OCR in this process.

Where there was a discrepancy between the code from Automatic Coding and the code from Query Resolution, we could display the image and the recognised text string for the response. We would then ask our staff doing the reprocessing to identify any errors in the text string.

90. This approach would allow us to differentiate between errors in the output of Automatic Coding which are the result of errors in Recognition, and other discrepancies between Automatic Coding and Query Resolution. In the process we could identify significant substitution effects.

Automatic Coding

91. The risk associated with Automatic Coding is that we will systematically code specific responses incorrectly.

92. Query Resolution is ideally suited to check the quality of Automatic Coding, since Query Resolution is set up to code responses to the same census questions. We will select a sample of responses coded in Automatic Coding to be recoded in Query Resolution. We will then compare the codes from the two processes for the same response, and determine which is the correct code where there is a discrepancy. By analysing those cases where the code from Automatic Coding is incorrect, we will identify systematic errors.

93. Another approach we are considering is to target the responses directly. We will store all responses. We could identify each unique response and rank them according to the frequency of that response. Since Automatic Coding will always assign the same code for a unique response, we would only need to check the code assigned to each unique response once. By simply working our way down the list from the most frequent responses, we could efficiently manage the quality of Automatic Coding.

Quality Management of Clerical Processes

94. Query Resolution and Expert Coding are the main clerical processes. We will follow the standard approach of reprocessing a sample of work to assess the quality of output. The way we use this information will be guided by the Total Quality Management philosophy which has a very good and well documented track record in the management of clerical processes.

95. The cornerstone of this philosophy is that deficiencies in the process itself cause most of the errors in output. We produce the training, systems, and procedures used in the process, therefore we are responsible for the quality of output, not our staff. Improving the process is the key to improving the quality of output.

96. Our people are our most valuable resource, as they are best placed to advise us on how to improve the process in which they work.

97. We will implement continuous quality improvement, using the following cycle;

i) measure quality
ii) identify the significant errors
iii) identify the root causes of these errors
iv) implement corrective action and measure quality once again.

98. We will use teams of processing staff, coding experts and managers to identify and resolve quality problems.

99. Our approach will focus on identifying and addressing the root causes of significant systematic errors in the clerical processes.

Process Control

100. Process control will be implemented to ensure that we send data through processes in the correct, logical order.

101. Each process will also produce information about changes applied to the data during that process. This information will provide an effective audit trail, allowing us to monitor changes in the data associated with each process, and to identify where data quality problems are introduced into the data.

V. Validation of Results

102. When we asked our customers to evaluate our 1991 performance, two of the most important areas of dissatisfaction were the significant errors in output and the fact that we had not achieved our timetable for the release of data.

103. In 1991 we did not validate the data until it was given to Census Output. The errors that were identified, and the subsequent correction of these errors, led to delays in the release of output. There were also a number of significant errors in the data that were not detected until after the release of output.

104. We are aiming to address both these areas by monitoring the data during processing and validating the results. A Validation Team will be set up specifically for this purpose. Their role will be part routine checking, part investigation and analysis of potential problems.

105. We will monitor data from the start of processing so that we can detect significant systematic errors as early as possible. By detecting errors early, we will have more time to address the sources of these errors and correct them. Identifying and addressing the sources of systematic errors early can also save considerable resources by reducing the need for, and scale of, reprocessing.

106. In some cases we may not be able to correct errors (for example if respondents are the source of the error), in which case we will be able to advise users as early as possible.

Routine Checks

107. The Validation Team will aggregate and tabulate data before implementing a series of checks. These checks will be performed at various geographic levels and at various stages of processing.

Check 1 - Totals

108. The “Totals” check will ensure that the table total is consistent with the data item we are looking at. For example, if a data item is applicable to persons then we will check that the total in the table is the same as the total number of persons in the population of that geographic area.

Check 2 – Not Applicable

109. The “Not Applicable” check will ensure that we have data for all records for which the data item is applicable. For example, if Occupation is only applicable for persons who are economically active then we will check that we have Occupation data for all persons who are economically active and do not have Occupation data for persons who are not economically active.

Check 3 – Intercensal Change

110. The “Intercensal Change” check will compare distributions in the table with distributions in the 1991 data for the same geographic area. Where the data from the 1991 Census is not comparable or was not collected in 1991, we will use alternative reference data if possible. This alternative reference data will include population estimates and a range of administrative data.

111. We will compare distributions (percentages of the applicable total), not counts.

112. We will set tolerance levels for the difference in distributions between 1991 and 2001. Where the 2001 distribution exceeds the tolerance levels for a given category, the Validation Team will investigate further.

113. For example, if we were validating the Sex variable, we would be looking at the distribution of males and females in the population. Let’s say the 1991 distribution for the area in question was 49% male and 51% female, and we set the tolerance for intercensal change at 1%. The data would pass the check if the 2001 distribution was 49.5% male / 50.5% female, but fail if the 2001 distribution was 50.5% male / 49.5% female.

114. When comparing data over a ten year period, there is no one level of intercensal change that is suitable for all data items and all geographic areas. As a general rule we would follow up any intercensal change of greater

than 5% at the county level (there were approximately 70 counties in Great Britain in 1991). For smaller areas there is likely to be much more variation, and we would generally set the tolerance at 10%. We will learn a lot about what level of intercensal change to expect for different variables from the results of the Census Dress Rehearsal in 1999.

115. The Intercensal Change check will also be implemented for a limited number of simple cross tabulations, particularly those included in standard output, and for particular sub-groups of the population.

116. We are investigating the potential of automating the Intercensal Change check. This would reduce the resources required to tabulate data and compare distributions, allowing us to spend more time investigating potential data quality problems.

Further Analysis

117. Where a data item fails any of the routine checks, the Validation Team will investigate further.

118. If the Totals or Not Applicable check fails, we clearly have a problem. We will use the information about changes to the data in each process to establish why we are either not coding the data item for some of the relevant population, or including (excluding) some of the population incorrectly as not applicable. We will then implement corrective action to correct the errors in the processed data and address the source of the errors so that subsequently processed data is not affected.

119. It is not so simple where the Intercensal Change check fails. This may or may not indicate a data quality problem. We will check that the data accurately reflects the responses on the forms. We will do this by selecting a sample of records and tracking the processing of the data item from Recognition through to the current process using the audit trail of changes to the data. Where we identify that an unexpected shift in the distribution for a data item is an error being introduced by one or more of our processes, then we will evaluate options for correcting the affected data and correcting the source of the errors in the process. Where the data accurately reflects the responses on the forms, we will analyse the change and prepare advice for our customers to expect this feature of the data.

120. For example, the 2001 Census form includes a Long-term Illness question that asks people whether they are limited in their daily activities or work by a long-term illness, health problem or disability. Respondents are asked to tick either “Yes” or “No”. Let’s say we compared the Yes / No distribution for 2001 with the distribution for the same data item in 1991, and found that there had been a shift from “Yes” to “No” which was greater than the tolerance we had set for Long-term Illness.

121. We would select a sample of records that had been coded to “No”. For these records we would check the images for the Long-term Illness responses on the forms, ensuring that the codes in the data after Recognition were

consistent with the tick responses on the forms. If the tick responses were being recognised incorrectly, we would correct the Recognition process and re-recognise all responses involved.

122. If we established that the tick responses were being recognised correctly, the next step would be to check the other processes where the data was being changed.

123. In the case of Long-term Illness, the other processes which change this data item are Edit and Imputation. Edit will resolve cases where a person ticks more than one tick-box. Imputation will impute a value for people who did not respond to the Long-term Illness question. We would test whether the distribution between “Yes” and “No” was changing significantly through Edit or Imputation. To do this, we would tabulate the data after Recognition and after Edit. We would then compare the distributions for Long-term Illness with that from the Intercensal Change check. If we found there was a significant difference between the distribution after Recognition and that after Edit or Imputation, we would evaluate options to reprocess or adjust data which had already been through these processes, and to address the source of the problem in the Edit or Imputation process.

124. If Edit and Imputation was not changing the distribution and we had captured the responses to the Long-term Illness question accurately from the forms, then we could not ask any more of our processing system. We might call this a clean bill of health for processing.

125. Responses to the Long-term Illness question are very much influenced by self-perception. Increased awareness of health issues in the community and government policy designed to increase access to employment for people with long-term illnesses may have influenced people’s perception of the extent to which their daily activities or access to employment are affected. This change may be part “real” (more activities and employment available to people with long-term illnesses) and part perception (people see other people with long-term illnesses doing activities and in jobs, and see their own illness or disability as less limiting). Our customers should be advised to consider these effects when making decisions based on the Long-term Illness data.

Data Quality Monitoring System

126. A Data Quality Monitoring System (DQMS) is being developed to allow the Validation Team to interrogate and analyse census data and information about changes to the data throughout processing. We are aiming for a “real time” environment as far as is practically possible.

127. To avoid affecting the performance of the processing systems, the DQMS will access a separate database, regularly loaded with the census data, information about changes to the data, and other information from the processing systems.

128. The Validation Team will be trained in data analysis techniques, and will use these techniques with the help of statistical analysis and tabulation

software linked to the DQMS. We are also evaluating the potential benefits of visualising the data geographically.

Expert Users

129. Our validation strategy includes seeking the help of expert users to assist us by checking data. Workplace data (where we geographically code people’s place of employment) is a good example of where expert users can supplement the expertise of the Validation Team.

130. The Workplace data is very prone to error, mainly because of the incomplete and incorrect responses we get for this question. Unfortunately the errors tend to be systematic and localised, resulting in large number of people being incorrectly coded to the wrong workplace “zone”. The errors can be very obvious where they involve an easily identifiable feature, such as a hospital, that is physically located in one “zone” but we all the hospital staff coded as working in another “zone”.

131. By enlisting the help of expert users of Workplace data to assist us in identifying systematic errors, we could benefit from their local knowledge or particular expertise (for example, with transport modelling). We would then adjust the data to correct or reduce the effect of these systematic errors.

Census Dress Rehearsal

132. The Census Dress Rehearsal will be conducted in 1999 to test our field and processing procedures and systems. This is an excellent opportunity to fully test our validation strategy, including the involvement of expert users. We will evaluate the quality of the data from the Dress Rehearsal, identify data quality problems, and fine-tune our systems and procedures to address these problems.

VI. Conclusion

133. Minimising the level of significant systematic error in output is a key objective for the 2001 Census of Population for the United Kingdom. We will achieve this objective with a balanced combination of quality control, quality management and validation of data before release to our customers.
