Information note 4: Data quality

On 5th May 2006 the responsibilities of the Office of the Deputy Prime Minister (ODPM) transferred to the Department for Communities and Local Government.

Department for Communities and Local Government
Eland House
Bressenden Place
London SW1E 5DU
Telephone: 020 7944 4400
Website: www.communities.gov.uk

Documents downloaded from the www.communities.gov.uk website are*Crown Copyright* unless otherwise stated, in which case copyright is assigned to *Queens Printer and Controller of Her Majestys Stationery Office.*

Copyright in the typographical arrangement rests with the Crown.

This publication, excluding logos, may be reproduced free of charge in any format or medium for research, private study or for internal circulation within an organisation. This is subject to it being reproduced accurately and not used in a misleading context. The material must be acknowledged as Crown copyright and the title of the publication specified.

Any other use of the contents of this publication would require a copyright licence. Please apply for a Click-Use Licence for core material at www.opsi.gov.uk/click-use/system/online/pLogin.asp or by writing to the Office of Public Sector Information, Information Policy Team, St Clements House, 2-16 Colegate, Norwich NR3 1BQ. Fax: 01603 723000 or e-mail: HMSOlicensing@cabinet-office.x.gsi.gov.uk.

This publication is only available online via the Communities and Local Government website: www.communities.gov.uk

Alternative formats under Disability Discrimination Act (DDA):if you require this publication in an alternative format please email alternativeformats@communities.gsi.gov.uk

Contents

1. Introduction

2. Briefing and fieldwork

3. Data checking and validation

4. Matching the data from the different surveys

5. Controlling for non response

6. Error levels in the survey

7. Statistical sampling error

8. Measurement error

9. Controlling surveyor variability

Appendix A

1. Introduction

1.1 This note describes the exhaustive set of procedures that were used to ensure that the data collected in the 1996 EHCS survey were as accurate as possible. Before any part of the survey was undertaken questionnaires and survey schedules were piloted, and all parts of the survey began with a briefing.

2. Briefing and fieldwork

Interview Survey

2.1 MORI managed and conducted the Interview Survey. The professional interviewers from MORI and NOP who were used on the survey each attended a one day briefing. These were held in different regions and were also attended by area managers and supervisors. Following on from the briefing each interviewer then conducted two dummy interviews before going out into the field.

Physical Survey

2.2 All 103 EHCS surveyors were either qualified building professionals, Environmental Health Officers with a private sector renewal background, or architects with renovation experience. The majority had previous EHCS experience, while the remainder were taken on recommendation.

2.3 Each surveyor attended a six day residential briefing which included lectures, discussions and practice surveys. All surveyors were provided with a comprehensive set of briefing manuals. Throughout the briefing surveyors were assisted by their Supervisor (1 supervisor per 10/11 surveyors) who also accompanied them for one day in the field (more if deemed necessary), and who monitored their progress and provided technical assistance.

Postal Survey

2.4 At the start of the fieldwork a lead officer was nominated by each local authority and housing association, who attended one of the one day briefings organised by MORI at which the procedures for completing the forms were explained. The forms were designed with explanatory notes accompanying each question, and a telephone help-line was provided to deal with queries about completion.

3. Data checking and validation

3.1 The first four completed forms of each interviewer were subject to a rigorous check by supervisors, and remaining forms were checked by clerical staff on receipt from the field for completeness. On data entry, the information from the interviewers was subjected to extensive editing, autocoding and autochecking to clean the data. All data were subjected to computer range checks and consistency checks. MORI were contracted to provide evidence of the quality of the data.

3.2 The first week's work of all surveyors was subjected to manual checks for completeness and consistency by survey supervisors. Once supervisors were satisfied with a surveyor's performance, forms were sent directly to MORI where they were double punched on to the computer and then run through a comprehensive validation routine to identify and rectify errors, inconsistencies, and implausibilities in the data.

3.3 Postal Survey forms were hand checked for completeness and consistency and, where necessary, local authorities and housing associations were asked to supply additional information. The data were then double punched onto the computer and a wide range of validation checks undertaken.

4. Matching the data from the different surveys

4.1 Exhaustive checks were undertaken to ensure that the address codes from each of the separate surveys that comprise the 1996 EHCS matched so that we could be sure that all the data related to the correct dwelling. For those dwellings which had also been surveyed in 1991 further checks were carried out, including interviewers using 1991 photographs to identify addresses to ensure that the same dwelling was visited in each year.

4.2 Checks were also undertaken to ensure that information that was obtainable from more than one source (tenure, dwelling type, dwelling age, for example) was consistent throughout the data set.

5. Controlling for non response

5.1 At various stages of the survey cases dropped out, or incomplete data were returned. It is important for the quality of the results that these cases are either demonstrated to be representative of the sample that remains, or adjustments are made to the data set to control for any identified 'non-response bias'.

5.2 Where non-response biases were identified they were dealt with at the grossing stage by applying compensatory weights to the cases that did respond. This process is described in more detail in Information Note 2 and in Appendix A to the main report.

5.3 Because of the 1996 EHC'S concentration on a core data set of cases with full surveys, the problem of missing data on individual variables has been minimalised. Where this does occur an assumption has been made at the tabulation stage that those dwellings/households with missing data should be distributed amongst the cells of the table in precisely the same way as dwellings for which data is complete.

6. Error levels in the survey

6.1 Any sample house condition survey will suffer from:

- sample error, associated with the use of a relatively small sample of dwellings to draw conclusions about the national dwelling stock and household totals;
- measurement error, due to inaccuracies in the individual measurements made by surveyors/interviewers because of difficulties inherent in the identification, interpretation and recording of what has been observed. This may be 'systematic' if a particular problem of observation is experienced by most, if not all, surveyors, or 'random' if individual surveyors are implicated differentially or inconsistently.

7. Statistical sampling error

7.1 Estimates of dwelling and household characteristics produced from a sample survey may differ from the true population figures because they are based on a survey rather than a complete census. This is known as sampling error, and it is important to know the extent of this error when interpreting the results.

7.2 The size of the sampling error depends on the size of the sample. In general, the smaller the sample size the larger the potential error. For example, estimates for dwellings in the private rented sector or converted flats will be subject to a larger sampling error than owner occupied dwellings or semi detached houses.

7.3 A way of taking account of sampling error is to calculate a confidence interval for an estimate. This is an interval within which it is fairly certain the true population figure lies. This section explains how 95 per cent confidence intervals can be calculated for the key survey estimates. The method suggested comes from standard statistical theory for large samples.

Confidence intervals for percentages

7.4 The 95% confidence interval for a percentage estimate, p, is given by the formula:

$$p +/- 1.96 * se(p)$$

where se(p) represents the standard error of the percentage and is calculated by:

$$se(p) = \sqrt{\frac{p(100-p)}{n}}$$ (n is the unweighted sample size)

7.5 Estimating standard errors for results based on a simple random sample, which has no stratification, are fairly straightforward. However, the sample for the EHCS is not a simple random one and so the standard errors could be corrected using a sample design factor. The design factor is calculated as the ratio of the standard error with a complex sample design to the standard error that would have been achieved with a simple random sample of the same size. Overall, the design effects for the EHCS were assumed to be small and so no adjustment has been made in the examples.

7.6 A 95 per cent confidence interval for a percentage may be calculated using Tables 1 and 2 given at Appendix A.

8. Measurement error

8.1 There are difficulties in assessing the condition of an individual dwelling or the characteristics of a household. The former mainly stem from the technical problems in the diagnosis and prognosis of defects found in dwellings. The difficulties are found particularly in the assessment of unfitness because of the subjective nature of the fitness standard, but also in the assessment of the state of repair. As a consequence, there is a high probability that two surveyors inspecting a given dwelling are likely to have different views on whether or not it is unfit and also on the extent and severity of disrepair.

8.2 The stock estimates of unfitness or disrepair rates based upon individual surveyor assessments are dependant on the > average performance = of all the surveyors. If a different surveying force had been used, then the estimate of the number of unfit properties would have been slightly different. So there is uncertainty or error associated with the estimate, and the greater the surveyor variability the greater is this error. It is therefore important to control this variability as much as possible and to understand the effect that any residual variability can have on the survey results.

9. Controlling surveyor variability

9.1 Experience has shown that surveyor variability cannot be completely eliminated or even reduced to an insignificant level, but precautions were taken during the 1996 EHCS to control its impact:

- by selecting a large sample of survey dwellings stratified to enhance the concentration of poorer condition properties and treating them in aggregate, to minimalise the impact of any deviant observations;
- by ensuring that the surveyors were provided with a rigorous and uniform briefing, which was backed up by survey manuals and supervision in the field;
- by re-employing surveyors from a mix of professional backgrounds who had taken part in the 1991 EHCS and who, as a group conformed to a normal distribution in their condition markings.

Understanding and measuring surveyor variability

9.2 The attempts to control for surveyor variability were tested in a number of ways:

- the number of unfit properties found by each surveyor was examined and compared with the 'expected' number;
- the number of unfit properties found by different professional groups of surveyors was examined to identify any bias which might result from different practices or common experience;
- the fitness judgements of surveyor's were compared with those of a different surveyor who 'called back' and undertook an additional survey at the same property as part of the quality control exercise.

Variation between individual surveyors

9.3 The proportion of properties found unfit by each individual surveyor was compared with the average for all surveyors. The problem with this test, however, is that the variability in unfitness rates could be due to real differences in surveyors' allocations of addresses, as well as to measurement error.

9.4 The test showed that the rates of unfitness recorded by surveyors formed a reasonably normal distribution. The mean proportion of properties found unfit was 8.4% (ungrossed), though there were some surveyors who found none unfit and others who recorded over 20%. The width of the distribution exceeded what would have been expected from sampling error alone, demonstrating that the dominant effect was surveyor variability. The normal shape of the distribution gave confidence that the use of averages over many surveyors would give a good estimate of the 'true' rate of unfitness.

Comparison between the fitness assessments of Environmental Health Officers and surveyors from other professions

9.5 Surveyors from a range of backgrounds are employed, notably from the professions of environmental health, building surveying and architecture, and from both the public and the private sector so that a range of judgements are made across all of the information obtained.

9.6 Of the 103 surveyors who took part in the survey, 47 were qualified Environmental Health Officers (the same proportion as in 1991) and the remainder were other building professionals. While application of the fitness standard is part of the daily routine of an EHO it is not for the other professions and so a rigorous common briefing is provided. This was, as far as possible, the same briefing as that which had been provided for the 1991 EHCS.

9.7 To establish whether the survey results were affected by using surveyors from different disciplines, comparisons have been made between the overall unfitness rate from dwellings surveyed by EHO's and from dwellings surveyed by other building professionals. The grossed results were as follows:

Environmental Health Officers 8.9%

Other building professions 6.4%

9.8 The EHOs appear to assess unfitness more severely, and there is a temptation to conclude from this that had the survey been carried out only by EHOs then the national rate of unfitness would be higher than the estimate of 7.5% presented in this report. However, an analysis of surveyor allocations shows that EHOs were more likely to be working in older metropolitan areas, where we would expect conditions to be worse. This is a result of the fact that public sector EHOs were more likely to be seconded to the survey from authorities with larger private sector renewal teams, and allocated addresses reasonably close to their homes. When we control for this there appears to be no significant difference between the groups, with the variation between surveyors within the groups being far greater than that between them.

Call-back survey

9.9 The Call-back Survey, which formed part of the quality control process, involved the revisiting and reporting on a sample of dwellings by second surveyors (different from the surveyors responsible for the first inspection, but drawn from the pool used for the main survey). The sample was drawn randomly from the main survey, but was stratified to include a higher proportion of dwellings in poor condition. Around 1,000 dwellings yielded sets of fitness assessments from both the main and the Call-back Surveys.

9.10 For the dwellings included in the Call-back sample the main survey yielded an unfitness rate of 29.4% while the Call-back Survey gave a rate of 22.4%. This suggests that there is unreliability in the measure of around 7%. However, this difference is likely to be an exaggeration because of the inconsistent ways in which the samples were selected for the main and Call-back Surveys. Dwellings were preferentially selected for call-back according to their poor condition as determined in the main survey, and in such a case it can be shown that the existence of assessment errors associated with the first inspection leads inevitably to an apparent improvement in average conditions by the second inspection. Had a disproportionate

number of good condition dwellings been selected for call-back then the opposite would have been the case.

9.11 Nevertheless this figure can be used in an exploratory way. A 7% error in an estimate of 29% is equivalent to an error of around 1.8% in the unfitness rate of 7.5% estimated for the national stock. The uncertainty associated with surveyor variability is therefore expected to be smaller than 1.8%, but almost certainly higher than the 0.4% attributed to sampling error as discussed in the section 7 above and in Appendix A. It is not possible to be more precise.

Appendix A

CALCULATING CONFIDENCE INTERVALS

1. A 95 per cent confidence interval for a percentage may be calculated using Tables 1 and 2 given below. The width of the confidence interval depends on the value of the estimated percentage and the sample size on which the percentage was based as shown in Table 1. For percentages based on the core sample, the sample size, n, is the unweighted sample total; that is 12,131 dwellings or 11,593 households. For estimates based on sub-samples, Table 2 lists the unweighted sample sizes for selected characteristics. The confidence interval can then be calculated by reading off the relevant figure from Table 1, with the estimated percentages shown as columns and the sample sizes shown as rows, and then adding and subtracting it from the estimated percentage.

Examples:

1 The estimated number of unfit dwellings is 1,522,000 or 7.5%. This percentage is based on the core sample of 12,131 dwellings. The corresponding number from table 1 is 0.4% giving a confidence interval of 7.1% to 7.9%[1].

2 The estimated percentage of unfit dwellings built before 1919 is 16.1%. This percentage is based on the sample of dwellings built before 1919, that is 3,257 (from Table 2). The corresponding number from table 1 is 1.4% giving a confidence interval of 14.7% to 17.5%.

Table 1: Look-up table for calculating 95 per cent confidence intervals for a percentage

Unweighted no. of dwellings	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%
12131	0.4	0.5	0.7	0.8	0.9	0.9	0.9	0.8	0.7	0.5	0.4
10000	0.4	0.6	0.8	0.9	1.0	1.0	1.0	0.9	0.8	0.6	0.4
9000	0.5	0.6	0.8	0.9	1.0	1.0	1.0	0.9	0.8	0.6	0.5
8000	0.5	0.7	0.9	1.0	1.1	1.1	1.1	1.0	0.9	0.7	0.5
7000	0.5	0.7	0.9	1.1	1.1	1.2	1.1	1.1	0.9	0.7	0.5
6000	0.6	0.8	1.0	1.2	1.2	1.3	1.2	1.2	1.0	0.8	0.6
5000	0.6	0.8	1.1	1.3	1.4	1.4	1.4	1.3	1.1	0.8	0.6
4000	0.7	0.9	1.2	1.4	1.5	1.5	1.5	1.4	1.2	0.9	0.7
3000	0.8	1.1	1.4	1.6	1.8	1.8	1.8	1.6	1.4	1.1	0.8
2000	1.0	1.3	1.8	2.0	2.1	2.2	2.1	2.0	1.8	1.3	1.0
1000	1.4	1.9	2.5	2.8	3.0	3.1	3.0	2.8	2.5	1.9	1.4
900	1.4	2.0	2.6	3.0	3.2	3.3	3.2	3.0	2.6	2.0	1.4
800	1.5	2.1	2.8	3.2	3.4	3.5	3.4	3.2	2.8	2.1	1.5
700	1.6	2.2	3.0	3.4	3.6	3.7	3.6	3.4	3.0	2.2	1.6
600	1.7	2.4	3.2	3.7	3.9	4.0	3.9	3.7	3.2	2.4	1.7
500	1.9	2.6	3.5	4.0	4.3	4.4	4.3	4.0	3.5	2.6	1.9
400	2.1	2.9	3.9	4.5	4.8	4.9	4.8	4.5	3.9	2.9	2.1
300	2.5	3.4	4.5	5.2	5.5	5.7	5.5	5.2	4.5	3.4	2.5

200	3.0	4.2	5.5	6.4	6.8	6.9	6.8	6.4	5.5	4.2	3.0
100	4.3	5.9	7.8	9.0	9.6	9.8	9.6	9.0	7.8	5.9	4.3

Table 2: Sample sizes of main variables for calculating confidence intervals

Variable	**No. of dwellings (weighted)** (thousands)	**Percentage (weighted)**	**Sample size (unweighted)**
All dwellings	20,371	100	12,131
Dwelling age	4,782	23.5	3,257
Pre 1919	3,900	19.1	2,671
1919-1944	4,255	20.9	2,851
1945-1964	7,433	36.5	3,352
1965+			
Dwelling type	6,189	30.4	4,394
Terraced	6,051	29.7	3,318
Semi-detached	4,199	20.6	1,307
Detached	3,013	14.8	2,630
Purpose-built flat	919	4.5	482
Converted-flat	16,439	80.7	9,019
All houses	3,932	19.3	3,112
All flats			
Tenure	20,371	100	12,131
All	13,928	68.4	5,988
Owner-occupied	2,032	10.0	906
Privately rented	3,470	17.0	3,437
Local authority (LA)	941	4.6	1,800
Housing association (HA)	19,573	100	11,593
Occupied dwellings	13,566	69.3	5,822
Owner-occupied	1,778	9.1	797

Privately rented	3,331	17.0	3,256
Local authority (LA)	898	4.6	1,718
Housing association (HA)	798	100	538
Vacant dwellings	616	77.2	125
Privately owned	139	17.4	181
LA owned	43	5.4	82
HA owned	19,573	96.1	11,593
All occupied	798	3.9	538
All Vacant			
Location	16,374	80.4	10,194
Urban	3,997	19.6	1,937
Rural			
Government office region	1,115	5.5	1,229
North East	2,042	10.0	1,395
Yorkshire and the Humber	2,336	11.5	1,124
North West	1,706	8.4	1,277
East Midlands	2,147	10.5	1,423
West Midlands	2,061	10.1	1,128
South West	2,234	11.0	1,053
East	3,198	15.7	1,002
South East	2,913	14.3	1,875
London	619	3.0	625
Merseyside			
Fitness of dwelling	1,522	7.5	975
Unfit	18,849	92.5	11,156
Not unfit			
Variable	**No. of**	**Percentage of**	**Sample size**

	households (weighted) (thousands)	**households (weighted)**	**(unweighted)**
All Households	19,675	100	11,593
Aged under 60	13,153	66.9	7,699
Aged 60+	6,522	33.1	3,894
Household type	6,730	34.2	3,406
Couple no dependent children	4,984	25.3	2,966
Couple with dependent children	1,253	6.4	1,129
Lone parents	1,317	6.7	864
Large adult households	2,271	11.5	1,217
One person under 60	3,120	15.9	2,011
One person aged 60+			
Household groups	776	6.0	482
Youngest households (aged 16-24)	848	6.5	590
Ethnic minorities (under 60 years)	1,129	8.6	990
Unemployed (under 60 years)	1,406	10.7	1,541
Economically inactive (under 60 years)	2,398	12.2	1,511
Elderly households (over 75 years)	2,703	13.7	1,887
Households with infants under 5 years	909	7.0	708
Sick/disabled households (under 60 years)			

[1] It should be noted that the sample design sought to minimise the error associated with the national estimate of unfitness. Thus, the confidence interval quoted in this example would be

narrower if a sample design factor was taken into account.