



Government Statistical Service
Methodology Series No. 7

Sample Design Options for an Integrated Household Survey

Dave Elliot & Jeremy Barton
ONS

May 1998

© Crown copyright 1998

ISBN 1 85774 269 9

ABOUT THE GSS METHODOLOGY SERIES

The aim of this series is to publish monographs with a substantial methodological interest written by people across the Government Statistical Service. Findings can be included provided that they illustrate methodology, but the series is not for findings *per se*.

Publication process

Contributors should send their submissions to the series editor at the address below. Before doing so, however, they should clear the work with their line manager. The submissions will then be subjected to peer review, external or internal, or both, as the series editor see fit. Authors can submit their work for peer reviewed journals as well.

The series is aimed at getting out results quickly and easily. It is the intention that authors will prepare the final documents themselves, on their own PCs. The style has therefore been kept simple. Style guidelines specifying the fonts, type size, margin settings, etc. will shortly be available from the series editor (at the address below). The series editor should need to do minimal editing.

Dissemination

Members of the GSS can obtain copies free of charge from the National Statistics Library, tel 0171-533 6257 or GTN 3042 6257, room DG/18 at the address below

Copies are also available for sale to customers outside the GSS. These can be obtained only from the National Statistics Sales Office, tel 0171-533 5678 room B1/06 at the address below.

Series editor

John Charlton
Office for National Statistics
Room DG2/08
1 Drummond Gate
London SW1V 2QQ
Tel 0171 533 6239

email john.charlton@ons.gov.uk

Navigate through this document by using Bookmarks, Thumbnails, or Links from the Contents listing below. There are more sub-categories in the bookmarks than there are in the Contents listing. Prevent the printing of these instructions by unchecking 'Annotations' in the Print dialogue box.

Select the HAND tool. Position it over the Contents list below and when an ARROW appears on the hand click the mouse button once to enlarge the text. Continue clicking the arrowed hand tool to advance down the list. When the HAND tool changes to a POINTED FINGER click the mouse button to link to the listed item.

Contents

	Page
1 Summary	5
2 Background	5
3 Broad Options	6
4 Stratification	9
5 Effects on Precision	11
5.1 Estimate from survey components	11
5.2 Estimates for core variables	13
5.3 Estimates for rare subgroups	13
6 Sample Rotation	14
7 Practical Considerations	16
8 Cost Implications	18
9 Appendices	
A Current sample designs from the core surveys	19
B Equivalent sample increases and standard error reductions	20
C Gains in effective sample size for selected variables	21
D Census variables used in options (C) and (D)	24
E Methods and estimates of precision	25
F Selecting the best general-purpose stratifiers	36
Publications in the GSS Methodology Series	52

This is a blank page

1. Summary

A number of possible sample designs for a combined survey, integrating four major continuous household surveys, are considered and it is concluded that worthwhile improvements in the precision of estimates from all four components: the Survey of English Housing, General Household Survey, Family Resources Survey and Family Expenditure Survey would be possible within current costs if a suitable design were to be adopted.

The advantages and disadvantages of a completely unclustered design and three different clustered designs based on different sized primary sampling units are discussed. An analysis of the effect of adopting a range of different stratifiers on key estimates from each of the four component surveys is reported and the best overall stratifiers identified. The effect of selecting a larger than normal fraction of areas is described. Finally estimates of the gains in precision from the combined effect of all these changes for a range of variables from each of the four component surveys are given.

From an analysis of the precision and likely cost implications, two of the clustered options can be readily eliminated. Consideration of the robustness of the designs to future changes in the composition and sample size required in the survey leads to the recommendation that a clustered design based on postcode sectors would be the best all-round option. A preliminary analysis of the additional costs of such a design indicates that, after the first year, the new design would provide better value for money than retaining the four separate surveys.

2. Background

In July 1995, Social Survey Division (SSD) of ONS carried out an investigation into the possibility of integrating a number of major continuous household surveys currently undertaken by ONS and SCPR on behalf of several different Government Departments. The aims were to standardise as much of the methodology and definitions as possible and then to combine the fieldwork and sample designs with a view to producing more reliable and consistent estimates from this group of surveys.

The central core of surveys would comprise DoE's Survey of English Housing (SEH), ONS's General Household Survey (GHS), the DSS's Family Resources Survey (FRS) and the ONS's Family Expenditure Survey (FES).

Following the presentation of the precision and cost implications and other information on the practical plausibility of integrating the fieldwork for the four surveys, the Inter-Departmental Steering Group charged with examining the case for such an integrated survey took the decision

not to proceed to a full implementation. In a note to the GSS Committee on Surveys of Persons and Households, they stated:

The field trials have shown that integration is feasible ... However, the commissioning departments also considered whether integration of outputs from the four surveys could be achieved independently of integration of the fieldwork. They concluded that it could ... Thus the decision on whether to pursue the option of integrated fieldwork reduced to the size of savings that could be achieved. These proved relatively low because of the need to maintain sample sizes at near their current levels in order to provide reliable estimates of small subgroups of the population.

Nonetheless the Steering Group agreed to sponsor the work necessary to complete the fieldwork feasibility experiments and to prepare methodological papers for publication. In addition there have been a number of spin-offs from the development work which should lead to improvement in training methods, sampling in multi-household addresses and harmonising questions and definitions across these and other GSS household surveys.

A detailed comparison on the current sample designs used by the FES and the FRS and a review of various options for bringing them into closer alignment was undertaken jointly by OPCS and SCPR in March 1994 but inclusion of the other two surveys allows a broader range of sample designs to be considered, and it is these which are discussed in this report.

An outline of the main features of the current sample designs for these four core surveys is given in Appendix A.

3. Broad Options

Four alternative sample design options are considered. In all four options it is assumed that the sampling frame would continue to be the PAF, that interviewers would work approximate monthly quotas of 20 household interviews and that each interviewer's quota would comprise a mixture of interviews for the component surveys.

The four designs discussed here are:

- a) An unclustered systematic sample of addresses, divided into areas for fieldwork convenience.
- b) A stratified two-stage sample of addresses within postcode sectors - essentially the current design on all four of the component surveys.
- c) A stratified two-stage sample of addresses within local authority districts, with each selected district being partitioned into 12 smaller areas, these being allocated to a particular month for fieldwork.

- d) A stratified two-stage sample of addresses within LFS Stint Areas¹, with each selected Stint being further partitioned into four smaller areas and allocated to months.

The rationale for considering these options is outlined below.

Variances of estimates made from surveys depend on the sample design, as well as the inherent variability of the variables in the survey population. It is customary to compare variances for a particular design with those from the simplest possible design in which the units are selected independently from the sampling frame with equal probabilities. This latter design is known as a simple random sample (srs) and the ratio of the variance of an estimator for a specific sample design to that for an srs is known as a design effect (deff). A simple random sample is not the most efficient possible and it can always be improved upon by stratification when suitable information exists and by selecting the units without replacement to avoid duplicate elements appearing in the sample.

When a continuous or multi-valued variable is available as a potential stratifier for a sample, a convenient way of achieving the stratification and selecting the sample without replacement is to sort the population elements by this variable and then select a systematic sample through the population. Single-stage systematic samples thus have design effects less than one and sometimes considerably less, when the variable concerned is significantly correlated with the stratifying variable. Such situations are comparatively rare in social survey research however.

The Postcode Address File (PAF) divides the country into c.8700 postcode sectors and thus the addresses can be ordered geographically prior to sample selection. The cost-effective design of survey samples for personal interview requires that interviewers do not have to travel long distances between addresses. This is usually achieved by clustering the interviewers' quotas of addresses by selecting the sample in stages. First a sample of areas is selected, then within the selected areas, a sample of addresses is selected. Estimates made from such designs have larger variances than those from single-stage designs and a careful weighing of the different parameters affecting costs and variances is important in any sample design process.

The parameter describing the effect of clustering the sample in this way is the intra-area correlation, ρ . Values of ρ depend on the size and nature of the areas used in a multi-stage design as well as the variable concerned. The effect on the variance of a survey estimate is largely determined by ρ and the average number of units selected within an area. The use of stratification in the initial selection of the areas can mitigate the effect of the clustering and

¹ These areas were created for the current LFS design and partition the whole of Great Britain into 1440 areas of approximately equal population.

even, on occasion, eliminate the extra variance. Usually, however, design effects for samples that employ both clustering and stratification are considerably greater than one.

For a variable with a large ρ for PAF sectors, because of the ordering by sector, a systematic sample through the PAF may eliminate a substantial part of the variance. If the sample is large enough, so that at least one unit is selected from each sector, the proportion of variance eliminated is at least ρ .

By merging a group of surveys, the overall sample size is increased and so an unclustered design becomes a possibility without incurring a large increase in interviewer travelling time.

Alternatively, by combining interviews from several surveys in a single interviewer's quota in a clustered sample, the number of units selected per area in any one of them is substantially reduced and hence the effect of clustering on the variances of estimates from a single survey can be drastically reduced and even, rarely, eliminated.

When estimates are based on characteristics of people or other units within households, design effects are often larger than for household characteristics because of the homogeneity within households. For example the relatively high design effect for the proportion of current smokers measured on the GHS is due in part to the area clustering effect but is mainly due to the homogeneity of this characteristic within households. For such variables there is less to be gained from an integrated survey design.

With option (a), no auxiliary information about addresses, beyond their location, is available from the PAF and no links are currently available between PAF addresses and other population sources so no further stratification is possible. Furthermore it is not appropriate to overlap the samples in successive years with this design so apart from the method of constructing the fieldwork areas and assigning them to months the description of this design requires no further elaboration.

The other options (b) - (d) all require a decision on the nature and form of the stratification to be used and this is discussed in Section 4, below. Section 5 provides estimates of the precision gains and losses of the four options above compared with the designs currently used on the component surveys, on the assumption that the achieved sample sizes do not change. Further details of the methods and results in this section are given in Appendix D. With the clustered designs it would be possible to retain all or a proportion of the areas from one year to the next (often described as rotation) and this is discussed in Section 6, below. Section 7 is concerned mainly with the robustness of the different designs and the practical implications of carrying them out. Finally Section 8 considers cost implications.

4. Stratification

Stratification is a simply-applied technique for improving the precision of survey estimates by ensuring that all "important" sections of the population are actually present in the sample. The optimum choice of variables used to stratify any survey will be those variables (available for the primary sampling units) which are most highly correlated with the variables of interest on the survey. For any survey, this choice is often complicated by the range of subjects under investigation - a stratification factor which is highly correlated with one survey variable may be poorly correlated with another. On the proposed integrated survey this problem is intensified, since the range of topics included would widen.

The PAF itself contains no additional information about addresses. Consequently, the way stratification is implemented in PAF samples in ONS is to link Census-derived information about areas to the sampling frame and to use this information when deciding which areas to select. Consequently stratification has most impact on those survey variables having large area-clustering effects.

Despite the wide range of topics included in the 4 component surveys the number of different stratifiers used is quite small, with some being used on more than one survey (see Appendix A).

The GHS and the SEH at present use the same stratifiers based on household tenure and socio-economic group, which reflect both surveys' roles in investigating household related outcomes. Likewise, the FRS and FES also share some stratifiers since both surveys cover topics related to income and expenditure. It may be that the best choice of stratifiers will be made from among those already in use.

If it were the case that the stratification factors had all been optimised for the main variables on the separate surveys, then a move to a single unified design would inevitably result in some loss of precision through the selection of a compromise set of stratifiers. In practice the only one of these surveys for which such an optimisation exercise has been carried through is the FRS. Previous work on the best choice of stratifiers for the FES suggests that a change of stratifiers would improve precision on important expenditure variables but this change has only recently been introduced so it has not been possible to measure the impact on precision in this study. The work on optimal stratifiers has now been extended to cover a range of variables on the GHS and SEH and the best compromise sought between the optimal solutions for the four component surveys.

Our method of comparing the impact of various alternative stratification factors was to fit main effects linear models to each of the survey variables measured at the postcode sector level. The goodness-of-fit was measured by the value for R^2 given by each model. A larger value of R^2 for a particular variable means that a greater proportion of the between-psu variance would be

removed through use of the model variables in defining strata². In addition to assessing the impact of a new set of stratification factors, the proposed stratifiers were compared with the existing stratifiers on each survey, to estimate the effect on precision should a new sample be adopted. Full details of these analyses are given in Appendix F.

As a result of this work, the stratifiers we propose for the integrated survey are as follows:

- i. Region (24 groups)
- ii. % Heads of household in SEGs 1-5 or 13 (4 - 6 bands)
- iii. % Households with no car (4 bands)
- iv. % Persons of pensionable age (2 or more bands)

With the exception of the tenure variables measured in the SEH, the proposed stratifiers perform on average as well or better than those currently in use on the component surveys and, thus, their introduction should improve precision on a range of important variables on the FES and the GHS while retaining current levels of precision on the FRS.

The situation with SEH estimates is more complicated. Clearly each of the tenure-based stratifiers would be the most effective for improving estimates of that particular tenure group but we cannot allow the tenure stratifiers to dominate the choice of stratifiers at the expense of the precision on other integrated survey estimates. The main stratifier on the current SEH is the percentage of private renters. This variable is weakly correlated with most other housing variables and with most variables on the other surveys and is therefore not a good all-round choice. The best housing-related stratifier is probably the percentage of local authority renters but this is a less effective stratifier on several SEH variables than the proposed set of stratifiers.

Each of the current surveys uses two or three stratifiers (in addition to a regional breakdown). The maximum effective number is limited by the total number of psus used in the design and the distribution across regions. With a larger number of psus the scope for using more stratifiers increases. If option (b) were adopted we should have a sufficiently large sample of areas to permit the use of a fifth stratifier and in that case we should recommend the use of the proportion of households renting from a Local Authority.

Previous work with the financial surveys has shown that the division of London into 4 quadrants improves precision more than the inner/outer London split currently employed on the GHS and SEH. The recent work has shown that adoption of the 4-way London split would make little difference on a range of GHS variables but does appear at first to reduce precision on a number of housing variables. However once the effect of our other proposed stratifiers has been built in this initial difference disappears and we therefore propose to use the 4-way split.

² This provided a good means of testing the various stratification options, although the value of R^2 is biased as an estimate of the population R^2 , as explained in Appendix E.

It is important to note that the gains in precision on SEH tenure variables as a result of other changes in the overall design will far outweigh any losses due to less effective stratification in all four options.

5. Effects on Precision

5.1 Estimates from survey components

The estimates of precision in this section are all given in terms of increases in effective sample size compared to that provided by the current separate designs. The effective sample size is the actual sample size divided by the design effect and is thus the sample size that would be needed in a simple random sample to produce the same precision as is produced by the design in question. Appendix B shows a number of equivalent ways of comparing the precision of the different designs.

The design on the SEH changed in April 1995, and it is this new design that is the basis of the SEH results. Changes to the FES stratifiers, recommended earlier, have now been implemented but these have not yet been in place for a full year and the impact on precision is not completely predictable so the comparison here still uses the old stratifiers.

The four factors affecting precision that have been used in estimating the gains from the integrated survey are:

- i. The intra-psu correlation, ρ . This parameter is both variable-specific and area-specific. It can be estimated for postal sectors from the data for each survey and for other areas it can be modelled using a combination of census and survey data. For an unclustered design, the intra-sector correlation estimates the size of the stratification effect.
- ii. The achieved number of survey-specific interviews per psu. Under option (b) the average values would be roughly 7 on the SEH, 3 on the GHS, 8 on the FRS and 2 on the FES.
- iii. The stratification factor. This is the proportion of the between-area variance that would be removed by stratification. This is variable-specific, stratifier-specific and, possibly, area-specific. It has been estimated for postal-sectors for several different sets of strata for a range of variables from each survey. For options (c) and (d) it is assumed that the factor is not area-specific which may have led to a slight over-statement of the precision gains from these options.
- iv. The proportion of all psus selected for the sample. The complement of this quantity is known as the finite population correction factor (for areas) and has the effect of dampening down the between-psu component of variance. With option (b) a sample of

39% of all postal sectors in Great Britain would be needed over the year to maintain the current sample sizes and with options (c) and (d) this rises to 55%.

All four options lead to improvements in precision for many of the key variables on the SEH, GHS and FES. On the FRS, option (c) reduces precision on a range of Census variables that one might expect to be correlated with income and so is not worth further consideration.

Table 1 summarises the increases in effective sample size for the full sample means of a range of important survey variables under the different options. The stratifiers used in each case are those recommended in Section 4. It should be noted that, for some of the variables, part of the increase in effective sample size is due to the change in stratifiers and that this change could be made without adopting an integrated design. However for other variables the choice of this compromise set of stratifiers reduces the effective sample size so it is the average net effect of all the changes to the design that is reported here. Detailed predictions for these variables for the whole sample and for a small number of subsamples are given in Appendix C. Also detailed estimates of the stratification and other effects are given in Appendix E.

It should be noted that the precision gains for small subgroups and for variables showing little or no clustering effect in the current designs will be negligible. Appendix C also shows predicted gains under options (a) and (b) in effective sample size for two subgroups of interest on the GHS: lone parent households, comprising 6% of the sample, and men in households with a manual head (20%). These are, predictably, smaller on most variables and were too small to be reliably measured in a number of cases.

Table 1. Average gains in Effective Sample Size for the Different Surveys

Survey	Option (a)	Option (b)	Option (c)	Option (d)
GHS	33%	28%	19%	32%
FES	31%	28%	9%	17%
SEH	131%	71%	10%	26%
FRS	16%	8%	-10%	5%

It is clear that options (a) and (b) significantly outperform options (c) and (d). As it was clear from an initial discussion with our Field Branch that any cost savings from these latter two options would probably be modest, they are not considered further. However option (a) significantly outperforms option (b) on the SEH and first indications suggested that Field costs for option (a) might increase by around 5% compared with option (b) so it cannot be quickly dismissed on cost grounds.

5.2 Estimates for core variables

In addition to these improvements in precision on the component-specific questions, the precision gains would be substantial on the core questions that would be asked with each component. For example, lone parent families could be identified from these core questions. In estimating the numbers of such families, the sample size would rise from 10,000 on the current GHS to 63,000 on the integrated survey. England would be over-represented in this sample so the data would need to be weighted to produce GB estimates and this would reduce precision slightly compared to an equal probability design. Under option (b), it is estimated that the effective sample size on the present GHS of 8,900 would rise to 57,400 on the integrated survey.

Similarly, since tenure is a core variable, under option (b) the effective sample size for the proportion of LA renters in England would rise from 10,200 on the current SEH to 27,200 on the integrated survey.

5.3 Estimates for rare subgroups

One of the attractive possibilities opened up by the integration of these four component surveys is that by pooling the samples one could identify a large enough sample of some rare subgroup of special interest to permit a separate analysis. For example suppose that some GHS client had a particular interest in the smoking and drinking habits of people from the ethnic minorities. In a normal year's GHS sample one would expect around 280 Asians and 170 Blacks. However in one year's integrated sample these numbers would rise to 1800 and 1100 respectively.

The simplest way to exploit these data would be to follow up those in the responding ethnic sample from all four component surveys who agreed to a recall. However this might jeopardise the response rate and an alternative would be to "swop" some ethnic minority households on the other surveys for non-ethnic households on the GHS. One would obviously not want to bias the results of the other surveys so one would want to retain some proportion of the ethnic minorities in order to weight them up to avoid this. The drawback of this approach is that the precision of all estimates on the donor surveys would suffer to some extent as a result of the necessary weighting. There would also be a rather larger effect on other estimates from the GHS, mainly as a result of the cut in the White sample on that survey.

Suppose in the above example, the GHS sample was boosted to 1200 Asians and 800 Blacks with a similar drop in the size of the White sample. As a result of this switch, one could expect the effective sample sizes on the other surveys to drop by 5% and for the GHS as a whole by 13%. Although one would not want to do this routinely, such a drop in precision might be tolerable in certain circumstances.

6. Sample Rotation

One method of improving the precision of estimates of year-on-year change is to retain all or part of the sample from one year to the next. These four surveys make heavy demands on the respondents so it is not considered feasible to retain the same households for a second interview but it is feasible to retain the same areas. For variables that show substantial systematic differences between areas (large clustering effects) but relatively little change over time, this "rotation" design can eliminate a significant part of the between area variance and so make a worthwhile improvement in precision.

The FRS is the only one of the four component surveys that currently uses such a design; it retains half the areas selected in each year for a second year. The SEH considered but rejected this design because of the impact it would have on the precision of survey estimates made by aggregating data for several years. Data from more than one year are also aggregated on both the FES and the GHS in a few applications but most of the analysis of these data sets is on the annual samples.

The gain in the precision (reduction in variance) of an estimated annual change by retaining areas in the sample from one year to the next is exactly matched by a loss in precision of estimates made by aggregating two years data. Losses are greater for aggregates of longer time periods though.

The change in precision depends principally on four factors - the stability of the variable whose change is being measured, the extent of the overlap from year to year, the sample size per area, and the variable-specific intra-area correlation.

The change in the effective sample size of retaining 50% of the sampled areas in an integrated design and selecting option (b) above is shown below for selected variables from the four component surveys. The first column shows the increase in effective sample size for measuring yearly change or alternatively the reduction in effective sample size when two years data are aggregated. The second column shows the corresponding reduction if three years data are aggregated.

Table 3. Changes in effective sample size in the integrated design if half the areas are retained

	Two years	Three years
SEH		
% Private renters	5%	7%
% LA renters	7%	24%
FES		
Total expenditure	5%	6%
Housing expenditure	5%	7%
FRS		
Average income	18%	25%
Average benefits	12%	17%
Average pension	12%	17%
GHS		
% smokers	5%	7%
% lone parents	1%	1%

The gains in precision for changes in FRS's income variables are much less in this new integrated design than in the current design but for these variables the loss in precision of estimates of annual change by dropping this feature in the new survey is almost exactly balanced by the improved precision arising from the other changes to the design. So in view of the reduction in precision of two and three year aggregates on the SEH, it is proposed that no area rotation be used with option (b).

7. Practical Considerations

With any of the designs we should need to devise a method of allocating the fieldwork to interviewers in manageable workloads. With the sector-based design, option (b), these would be the sector samples themselves. For the unclustered systematic sample the areas containing 20 responding households would on average be 2.5 times the size of a postcode sector. The most convenient way of forming such areas would be as groups of whole postcode sectors but there is no ready-made grouping available so this would need to be done with maps (or a mapping package) taking account of the varying population sizes in the sectors.

Our experience with the LFS, where we had to undertake a similar process, suggests that this is a very time-consuming task and the classification needs to be refined over the first year to make it practicable. An important feature of the LFS fieldwork design that makes it considerably more cost-effective is that the individual fieldwork quotas are formed into roughly contiguous groups of 13, each of which is allocated to a particular week, and the whole group is worked by a single interviewer. The comparable design for the integrated survey under option (a) would require the 3168 quota areas to be divided into 264 larger areas to be worked over the whole year.

The total number of households to be interviewed over a year on the core surveys is comparable to the number of household interviews on the LFS in one quarter. However the core surveys require more intensive fieldwork than the LFS which requires only one call at each address once the co-operation of the household has been gained. Consequently one would expect an interviewer to have to make rather more visits to the area per sampled case than on the LFS and for there to be more mopping up of late respondents outside the monthly field periods with consequently greater travel between areas on an integrated survey. This makes the final fieldwork costs of such a design difficult to predict, although we do know that on the LFS fieldwork costs increased by around 5% compared with a sector-based design.

The main concern with design (a) however is not its immediate impact on costs but its robustness in the face of future changes to the content and particularly the size of the integrated survey. If such a design were employed for the four core surveys and one of these later dropped out or had its sample size significantly reduced, it is difficult to see how we could prevent costs rising on the remaining component surveys. Furthermore if other surveys were to be incorporated at a later date we might have to redesign the fieldwork areas completely with some inevitable disruption. The alternative that we under-employ the interviewers in the initial design by setting monthly quotas less than 20 is no more attractive since it would raise initial costs and risk increasing interviewer turnover if we could not keep them fully occupied.

With the sector-based option (b) there would be no such problems since we would increase or reduce the number of sectors in response to changes in the composition of the survey.

Although precision would drop on the remaining surveys if one of them withdrew from the sample, there would be no effect on costs and as new surveys joined, the precision of the existing components would improve.

The SEH is obviously confined to England. With option (b), to minimise the number of Scottish and Welsh interviewers we could scale up numbers of addresses selected in each area for the 3 component surveys covering these countries and scale down the number of areas in proportion so that we end up with a single fixed quota size in all areas.

A design with close to the current overall set sample size of 94,944 PAF addresses could be achieved with a sample of 2850 sectors in England and 318 in Wales and Scotland with set quotas of 30 addresses. PAF ineligibility and non-response would reduce this to around 20 responding households. There would need to be some variation from the current set sample sizes on the four component surveys, in order to fix the component survey quota sizes as whole numbers. For example, these could be set as follows.

For England:

GHS	4 per quota
FES	3 per quota in 50% of the sectors 4 per quota in 50% of the sectors
FRS	12 per quota in 50% of the sectors 13 per quota in 50% of the sectors
SEH	10 per quota

For Wales and Scotland:

GHS	6 per quota
FES	5 per quota
FRS	19 per quota

With no cut in the sample size, these would produce set sample sizes for the 4 component surveys as follows:

Survey	Old	New
GHS	13248	13308
FES	11424	11565
FRS	42048	41667
SEH	28224	28500

8. Cost Implications

Under option (b) we believe there would be little or no change in survey running costs. However there would be a substantial initial set-up cost, mainly to cover the initial briefing of interviewers on all the component surveys. The additional cost of adopting the unclustered design (a) is more difficult to predict. Current indications are that fieldwork costs (interviewers fees and expenses) would rise by around 5%.

Appendix A. Current Sample Designs for the Core Surveys

	SEH	GHS	FRS	FES
Coverage	England	GB	GB	UK
Set Sample	28224	13248	42924	11424
Achieved Sample	20000	10000	25000	7300(GB)
Areas sampled	1008	576(GB)	1752	672(GB)
Stratifiers	Region-1	Region-1	Region-2	Region-2 ³
	% Private Renters	% Private Renters	% SEG 1-5, 13	Population Density
	% LA Renters	% LA Renters	% Unemployed	% Owner-occupiers
	% SEG 1-5, 13	% SEG 1-5, 13	% Owner-occupiers	% Private Renters
Period represented	quarter	year	month	quarter
Other features			50% of areas retained each year	

Region-1 divides Great Britain into 22 regions, comprising whole⁴ local authority districts and boroughs with London being divided into two, inner and outer. Region-2 is identical except that London is divided into 4 approximate quadrants.

³An earlier report recommended that these be changed to Region 2, SEG, % households who own a car, and male unemployment rate. The first three stratifiers were implemented in 1995/96.

⁴Actually approximations to LA districts comprising whole postcode sectors

Appendix B. Equivalent Sample Increases and Standard Error Reductions

Increase in Effective Sample	Reduction in Variance	Reduction in S.E.
10%	9%	5%
20%	7%	9%
30%	23%	12%
40%	29%	15%
50%	33%	18%
60%	38%	21%
70%	41%	23%
80%	44%	25%
90%	47%	27%
100%	50%	29%

If x is the increase in effective sample size, the reduction in variance is calculated as

$$\frac{x}{I + x}$$

This can also be interpreted as the reduction in the sample size with the integrated survey that would provide the same precision as the current separate survey.

The reduction in the standard error is calculated as

$$1 - \sqrt{\frac{I}{I+x}}$$

Appendix C. Gains in effective sample size for selected variables

Table C1. Gains in Effective Sample Size for Selected SEH Variables

	Option (a)	Option (b)	Option (c)	Option (d)
% Housing. Association renters	227%	114%	6%	31%
% Private renters	33%	17%	9%	30%
% LA renters	141%	58%	6%	29%
% Owner occupiers	96%	51%	1%	29%
% Living in terraces	187%	100%	-	-
% Living in flats	144%	82%	-	-
% Detached	143%	92%	-	-
% Semi-detached	162%	88%	-	-
Rooms per person	45%	35%	-	-

Table C2. Gains in Effective Sample Size for Selected FES Variables

	Option (a)	Option (b)	Option (c)	Option (d)
All expenditure	59%	54%	-	-
Housing expenditure	65%	57%	-	-
Fuel expenditure	91%	80%	-	-
Food expenditure	36%	32%	-	-
Alcohol expenditure	32%	29%	-	-
Tobacco expenditure	26%	24%	-	-
Clothing expenditure	13%	12%	-	-
Household goods exp.	10%	9%	-	-
Household services exp.	11%	11%	-	-
Personal goods & services.	1%	1%	-	-
Motoring expenditure	9%	8%	-	-
Fares expenditure	2%	2%	-	-
Leisure expenditure	8%	7%	-	-

Table C3. Gains in Effective Sample Size for Selected FRS Variables

	Option (a)	Option (b)	Option (c)	Option (d)
Gross income	25%	14%	-	-
Earnings	28%	16%	-	-
Self-employed income	3%	2%	-	-
Benefit income	24%	12%	-	-
Pension income	21%	15%	-	-
% Family credit recipients	1%	1%	-	-
% Income support recipients	10%	5%	-	-
% Housing benefit recipients	25%	14%	-	-
% One parent benefit recipients	29%	4%	-	-
% Unemployment ben recipients	21%	9%	-	-

Table C4. Gains in Effective Sample Size for Selected GHS Variables

	Option (a)	Option (b)	Option (c)	Option (d)
% persons: visited GP	7%	6%	-	-
% adults: frequent light drinkers	19%	18%	-	-
% persons: limiting long-standing illness	13%	12%	3%	7%
% adults: current smoker	19%	17%	-	-
% adult women: working	28%	25%	-	-
% adult men: working	9%	9%	-	-
Household size	30%	28%	-	-
% Lone parents	13%	13%	8%	11%
% HOH: married	47%	40%	-	-
% HOH: white	142%	110%	23%	79%
Length of residence	33%	28%	-	-

Table C5. Gains in Effective Sample Size for Subsamples for GHS Variables

	Option (a)	Option (b)
Lone Parent Households		
% Adults: A levels	0	0
% Persons: GP in last 2 weeks	32%	27%
% Adults: frequent light drinkers	0	0
% Persons: limiting long standing illness	0	0
% Adults: current smokers	10%	9%
% Adult women: working	14%	12%
% Adult men: working	0	0
Average household size	2%	2%
% Heads of household: married	18%	16%
% Heads of household: white	81%	70%
Average length of residence	68%	60%
Men in Households with a Manual Head		
% with A levels	0	0
% visited GP in last 2 weeks	0	0
% who are frequent light drinkers	7%	6%
% with limiting long standing illness	0	0
% who are current smokers	17%	15%
% who are working	7%	6%

Appendix D. Census variables used in options (c) and (d)

% Households renting from a housing association	(SEH, GHS, FES)
% Renting from a local authority or new town	(SEH, GHS,FRS, FES)
% Renting privately, furnished	(SEH, GHS,FRS, FES)
% Renting privately, unfurnished	(SEH, GHS,FRS, FES)
% Renting privately, with job	(SEH, GHS,FRS, FES)
% Owner-occupiers	(SEH, GHS, FES)
% Households in non-permanent accommodation	(SEH, GHS,FRS)
% Households in shared accommodation	(SEH, GHS,FRS)
% Households with > 1.5 persons per room	(SEH, GHS,FRS)
% HOH in SEG 1-15 or 13	(GHS,FRS, FES)
% Households with no car	(GHS,FRS, FES)
% Households with 3 or more cars	(GHS, FES)
% Households with a single parent	(SEH, GHS,FRS, FES)
% Households with a black HOH	(GHS,FRS)
% Households with an Asian HOH	(GHS)
% Persons born in NCW	(GHS)
% Persons with limiting long-standing illness	(GHS)
% Persons of pensionable age	(GHS)
% Persons aged 75 and over	(GHS)
% Persons aged under 16	(GHS)

Appendix E. Methods and estimates of precision

a. Methods used to derive the main precision predictions

The basic model used to estimate design effects for household-level variables here is:

$$deff = 1 + \rho [\bar{b}(1 - f_a)(1 - R^2) - 1] \quad (1)$$

where \bar{b} is the average number of responding cases per area (psu); ρ is the intra-psu correlation for the variable in question; f_a is the sampling fraction for areas; and R^2 is the squared correlation between the area means for the variable in question and the stratifier values (see below).

The use of f_a as a sampling fraction here is an approximation to the true sampling fractions which vary by area in the pps designs used in all 4 of these surveys. Wolter⁵ suggests an alternative approximation for this case given by $\bar{\pi}$, where π_i is the inclusion probability for area i . This appears to give a larger value whenever the psu sizes vary so the approximation used here should be conservative.

The method of estimating R^2 for the current stratifiers is to calculate r^2 , the squared correlation between variable means and stratum indicators for the sampled areas. These are calculated using a main effects, non-nested regression model for region and the first two stratification factors, as grouped in the current design. The third factor is fitted as a linear term to reflect its division in practice into however many implicit strata remain in each region x first stratifier x second stratifier combination. The omission of the interaction terms may lead to a slight understatement of the true impact of stratification but this method is preferred to a direct estimate of the collapsed strata effect since it retains a larger number of degrees of freedom and because it can be straightforwardly extended to provide a prediction of the effect of stratification when different stratifiers are used..

However r^2 will always be an underestimate of R^2 because of the sub-sampling of households within the selected areas. The covariance is unaffected by this but the variance between area means will be inflated by the within-area variance. If

$$R^2 = \alpha r^2 \quad (2)$$

then

⁵Wolter (1985) *An Introduction to Variance Estimation*. Springer-Verlag

$$\alpha = \frac{1 + (\bar{b} - 1)\rho}{\bar{b}\rho} \quad (3)$$

However ρ is unknown so $deff$ is estimated directly for the current stratifiers (by EPSILON). Then equation (1) without the area-level fpcf (see below), (2), and (3) are solved jointly to give

$$\alpha = \frac{(\bar{b} - 1)deff}{\bar{b}(deff - (1 - r^2))} \quad (4)$$

and

$$\hat{\rho} = \frac{deff - (1 - r^2)}{(\bar{b} - 1)(1 - r^2)} \quad (5)$$

For a new set of stratifiers, r^2 is calculated from a similar main effects model and then multiplied by the same value of α to give the new predicted R^2 .

Because the directly estimated design effects (from EPSILON) contain no adjustment for the area-level finite population correction factor (fpcf), the model was fitted to the existing $deff$ s with the fpcf term omitted and adjusted $deff$ s were constructed and used in all the comparisons. It is these adjusted $deff$ s that are reported below.

The design effects for the systematic sample are estimated as $(1 - \rho_s)$, where ρ_s is the intra-sector correlation, estimated as above. The FES has the smallest achieved sample of 7300 households in Great Britain and these would be distributed among the 8200 (grouped) postal sectors used on the current survey samples if a systematic sample were selected from the sorted PAF so that most of the sectors would be represented in the sample. With the larger surveys, all or virtually all of the sectors would be included. So the sectors can be treated approximately as strata and hence that component of the variance associated with differences between sectors would be removed in a systematic sample.

The method used to estimate design effects for individual-level variables (and also appropriate to any other unit which is clustered within households) is similar. In this case however a component of the current design effect will be due to the clustering within households and this component will not be affected by any of the alternative designs being proposed here.

The model used here is

$$deff = \frac{n}{h} \left[\rho \bar{b}(1 - f_a)(1 - R^2) + \frac{v_h}{\sigma^2} \right] \quad (6)$$

where n is the number of individuals in the sample, h is the number of households in the sample, σ^2 is the unit variance between individuals in the population and v_h is the between household (within area) component of the individual level variance. \bar{b} as before is the average number of households interviewed per area.

v_h/h is estimated directly from EPSILON by defining an area as a "stratum", a household as a "psu" and using the differences from the mean variance formula for an individual variable.

If

$$deff_h = \frac{n}{h} \left[\frac{v_h}{\sigma^2} \right] \quad (7)$$

and analogously to the household variable case

$$R^2 = \beta r^2 \quad (8)$$

Then

$$\beta = \frac{n\bar{b}\rho + h deff_h}{n\bar{b}\rho} \quad (9)$$

As before (6), (8) and (9) must be solved jointly to give

$$\beta = \frac{deff}{deff - (1 - r^2) deff_h} \quad (10)$$

and

$$\hat{\rho} = \frac{h(deff - (1 - r^2) deff_h)}{n\bar{b}(1 - r^2)} \quad (11)$$

Otherwise the method of predicting deffs for the new designs is identical to the household variable case.

The design effects for Local Authority Districts and for LFS Stint Areas were estimated from small area Census data using model (1). Thus with one or two exceptions, it was not possible to use variables specific to the individual surveys for these estimates. Instead a range of Census variables was selected and used to provide estimates for the four component surveys both for these larger areas and for postcode sectors to obtain relative effects.

The census small area data has so far only been extracted for EDs, Wards, LADs, Counties and GHS Regions. For each variable the values of ρ for each area type were plotted against average size on a log-log scale. Most of these relationships were approximately linear, so linear interpolation was used to provide an approximation to ρ for postcode sectors. It was these values that were used in the above comparisons between LAD-based designs and the current designs. The same interpolation method was used to estimate ρ for LFS Stint Areas.

Unfortunately the effects of stratification could only be modelled very approximately. These had been evaluated for the specific strata used on the GHS (which are based on SEG, % Private renters and % LA renters) for an ED-based sample. It was assumed that the same proportion of the between-area variance would be removed by stratification in samples based on the larger areas.

For the GHS estimates some person-level Census variables were used but in this case no allowance was made for clustering within households as this component of the variance cannot be estimated from the Census data.

b. Checks on the estimates

The Census data provides an independent check on ρ (for wards) for a few variables that are duplicated on these four component surveys. Obviously there is no guarantee that the variables are defined in exactly the same way nor that there have been no relevant changes over time between the Census and the survey, but one would expect the two to be broadly consistent.

The variables that can be checked in this way are some of the tenure categories: private renting, renting from a local authority or new town and owner-occupation estimated on the SEH. The single parent variable on the GHS is rather more problematic since it is not entirely clear that the population bases are sufficiently comparable and because the Census results have only been derived for England and Wales, while the GHS estimates apply to the whole of Great Britain.

Estimates of ρ from these two sources are shown in Table E1.

Table E1. Estimates of ρ from different sources

Variable	Survey	Census
Private renter	4.1%	6.6%
LA renter	18.1%	16.3%
Owner-occupier	14.4%	13.8%
Single parent	2.0%	1.6%

The agreement is not particularly close, but the two estimates are broadly in line. In three of the four variables, the survey estimates of ρ for postcode sectors are slightly larger than the Census values of ρ for wards, whereas one would have expected the true values for the slightly larger sectors to be lower than those for wards. This may be because the stratification model is under-estimating the true effect of stratification so the estimated ρ is being increased a little to compensate. Alternatively this may reflect either differences between the survey and Census variables or imprecision in the estimates.

In the case of the FRS income components and benefit receipt data, ρ was also estimated directly from the variance between households within areas, rather than from equations (2)-(5) above. This therefore avoids the modelling of stratification effects. The results are given in Table E2

Table E2. Comparison of Direct and Model-based Estimates of ρ

Variable	Direct	Model
Gross income	6.7%	8.2%
Earnings	5.5%	6.7%
Self-employed income	1.1%	1.0%
Benefit income	5.3%	6.0%
Pension income	4.3%	3.4%
% Family credit recipients	0.6%	0.8%
% Income support recipients	6.2%	6.6%
% Housing benefit recipients	9.3%	12.7%
% One parent benefit recipients	1.4%	1.6%
% Unemployment benefit recipients	0.4%	0.3%

These are in moderately close alignment. Although the majority of the model-based estimates of ρ are larger than the corresponding direct estimates, the differences are mainly small and do not suggest that ρ has been consistently over-estimated by the approach taken. In the case of the individual-level GHS variables the effects of the current stratification were also estimated directly from the variance components as well as from the regression model.

Table E3. Comparison of Direct and Model-based Estimates of Stratification Effect

Variable	Direct	Model
Adults with A levels	69%	73%
Consulted GP in last 2 weeks	37%	46%
Frequent light drinkers	78%	76%
Limiting long-standing illness	63%	49%
Current cigarette smokers	70%	72%
Men currently working	21%	48%
Women currently working	73%	78%

Once again, the majority of the model-based estimates of R^2 are higher than the direct estimates. In this case the direct estimates will always under-estimate the true stratification effects to some extent since they cannot take full account of the final stratifier.

c. Detailed estimates for the four component surveys

The four Tables below show current adjusted design effects, ρ , R^2 for both the current and proposed stratifiers and predicted design effects for options (a) (systematic) and (b) (sector-clusters) for each variable included in the analysis.

Table E4. GHS $f_a = .071$, $a=576$, $h=10,000$, \bar{b} (now)=17.4, \bar{b} (IHS(b))=3.2

Variable	ρ	Current deff	Current R ²	New R ²	New deff(b)	New deff(a)
<i>Household level</i>						
Household size	0.026	1.26	0.32	0.65	0.99	0.97
Lone parent	0.020	1.11	0.60	0.67	0.99	0.98
Married HOH	0.070	1.39	0.58	0.60	0.98	0.93
Non-white HOH	0.125	2.11	0.39	0.43	1.01	0.87
Length of residence	0.033	1.29	0.39	0.39	1.01	0.97
<i>Person level</i>						
A levels +	0.065	1.67	0.73	0.70	1.28	1.23
Consulted GP in last two weeks	0.004	1.22	0.46	0.48	1.14	1.13
Frequent light drinkers	0.035	1.53	0.76	0.86	1.30	1.28
Limiting long-standing illness	0.008	1.43	0.49	0.64	1.28	1.27
Current smoker	0.031	1.54	0.72	0.77	1.31	1.29
Women currently working	0.035	1.33	0.48	0.58	1.07	1.04
Men currently working	0.029	1.12	0.78	0.89	1.03	1.02

Table E5. FES $f_a = .083$, $a=672$, $n=7,200$, \bar{b} (now)=10.8, \bar{b} (IHS(b))=2.3

Variable	ρ	Current deff	Current R ²	New R ²	New deff(b)	New deff(a)
Total expenditure	0.096	1.40	0.49	0.72	0.94	0.90
Housing	0.105	1.43	0.49	0.69	0.94	0.89
Food	0.065	1.24	0.52	0.64	0.97	0.93
Motoring	0.025	1.05	0.69	0.81	0.98	0.98
Fuel	0.133	1.60	na	na	0.92	0.87
Alcohol	0.051	1.23	na	na	0.97	0.95
Tobacco	0.043	1.19	na	na	0.98	0.96
Clothing	0.022	1.10	na	na	0.99	0.98
Household goods	0.017	1.07	na	na	0.99	0.98
Household services	0.019	1.09	na	na	0.99	0.98
Personal goods and services	0.001	1.00	na	na	1.00	1.00
Fares	0.003	1.01	na	na	1.00	1.00
Leisure goods	0.013	1.06	na	na	0.99	0.99

Although the effect of stratification was not estimated directly for a number of these variables, estimates of ρ and the design effects were constructed using the average effects for the first four variables.

Table E6. SEH $f_a = .148$, $a=1008$, $n=20,000$, \bar{b} (now)=19.8, \bar{b} (IHS(b))=7.1

Variable	ρ	Current deff	Current R^2	New R^2	New deff(b)	New deff(a)
HA renter	0.139	2.82	0.17	0.17	1.35	0.86
Private renter	0.041	1.27	0.55	0.20	1.10	0.96
LA renter	0.181	1.93	0.64	0.39	1.28	0.82
Owner-occupier	0.144	1.65	0.67	0.52	1.14	0.86
Terraced house	0.144	2.45	0.35	0.41	1.22	0.86
Flat	0.201	1.94	0.66	0.69	1.06	0.80
Detached house	0.194	1.97	0.65	0.70	1.05	0.81
Semi-detached house	0.128	2.27	0.36	0.36	1.22	0.87
Rooms per person	0.044	1.38	0.42	0.63	1.02	0.96

Note that although these estimates are based on data from the 1994 survey, before the increase in the number of areas from 780 to 1008, the "current" design effect has been adjusted to take this into account. However it was not possible to take account of the change in the stratifiers that was introduced at the same time so that the "current" R^2 may be understated. No tenure variable was included in estimating the new R^2 , so the new deffs under option (b) may similarly be overstated on some variables.

Table E7. FRS $f_a = .217$, $a=1752$, $n=26,000$, \bar{b} (now)=15.0, \bar{b} (IHS(b))=8.4

Variable	ρ	Current deff	Current R ²	New R ²	New deff(b)	New deff(a)
Gross income	0.082	1.15	0.75	0.76	1.01	0.92
Earnings	0.067	1.19	0.66	0.68	1.03	0.93
Self-employed income	0.010	1.02	0.74	0.80	1.00	0.99
Benefit income	0.060	1.17	0.66	0.63	1.04	0.94
Pension income	0.034	1.17	0.47	0.70	1.01	0.97
% Family credit recipients	0.008	1.02	0.66	0.89	1.00	0.99
% Income support recipients	0.066	1.13	0.75	0.69	1.03	0.93
% Housing benefit recipients	0.127	1.13	0.83	0.66	1.08	0.87
% One parent benefit recipients	0.016	1.08	0.49	0.49	1.02	0.98
% Unemployment benefit recipients	0.003	1.01	0.72	0.84	1.00	1.00

d. Rotation effects

Rotation effects were estimated using the variance formula given by Holt and Farver⁶ for a design which retains half the areas from one year to the next. i.e.

$$\text{Var}(\bar{y}_t - \bar{y}_{t-1}) = 2 \frac{\sigma^2}{n} \left[1 + \left[\bar{b} \left(1 - \frac{c}{2} \right) - 1 \right] \rho \right] \quad (12)$$

where c is the correlation between area means one year apart (taken as 0.9) and ρ is the intra-area correlation. The variance of the two-year mean is given by:

$$\text{Var}\left(\frac{\bar{y}_t + \bar{y}_{t-1}}{2}\right) = \frac{\sigma^2}{2n} \left[1 + \left[\bar{b} \left(1 + \frac{c}{2} \right) - 1 \right] \rho \right] \quad (13)$$

⁶Holt & Farver(1992) The use of composite estimators with two-stage repeated sample designs. *Journal of Official Statistics*, 8, 405-416

This formula ignores the effect of stratification and sampling without replacement, so a combined measure, ρ' , was used instead of ρ . This is defined as:

$$\rho' = (1 - f_a)(1 - R^2)\rho \quad (14)$$

In the case of a three-year average, (13) has been generalised to:

$$\text{Var}\left(\frac{\bar{y}_t + \bar{y}_{t-1} + \bar{y}_{t-2}}{3}\right) = \frac{\sigma^2}{3n} \left[1 + \left[\bar{b} \left(1 + \frac{2c}{3}\right) - 1 \right] \rho \right] \quad (15)$$

In the case of an individual-level variable, (12) is modified to

$$\text{Var}(\bar{y}_t - \bar{y}_{t-1}) = 2 \frac{\sigma^2}{n} \left[\text{deff}_h + \frac{n}{h} \bar{b} \left(1 - \frac{c}{2}\right) \rho' \right] \quad (16)$$

where deff_h is defined in (7).

Appendix F. Selecting the best general-purpose stratifiers

a. Recent work on stratification of the FES and FRS

Recent work has been carried out independently on finding the optimum stratifiers for the two financial surveys, the FRS⁷ and the FES⁸. The results of these analyses were comparable, as would be expected considering the similarity in subject matter of the two surveys, although the variables analysed were different. The FRS analysis used gross household income and net housing costs, whilst the analysis for the FES used weekly household expenditure, net housing expenditure, motoring expenditure, food expenditure and, once again, gross household income. The stratifiers selected for each survey are now in use. The analyses will not be repeated, but a selection of income variables were tested against the FES, the FRS and the proposed IHS stratifiers.

b. The Census stratifying variables

The variables used to stratify surveys were derived from the 1991 Census and aggregated to postcode sector level. The variable descriptions with their corresponding short names are shown in Table F1. These short names will be used below for brevity. Only the GHS and FES regional breakdowns were analysed, because previous analyses on FRS and FES data showed the standard and standard/metropolitan regions to be poor predictors of those surveys' variables.

⁷ Bruce, S. (1993) Selecting Stratifiers for the Family Resources Survey. *Survey Methodology Bulletin*, 32, 20-25. OPCS
⁸ Barton, J. (1996) Selecting Stratifiers for the Family Expenditure Survey. *Survey Methodology Bulletin*, 39, 21-26. ONS

Table F1. Variables available for stratification

Short name	Variable description
Standard Region	Standard regional breakdown - 11 bands
Met Region	Standard and metropolitan regions - 16 bands
GHS Region	GHS regional breakdown - 22 bands
FES Region	FES regional breakdown - 24 bands
%Owners	% Owner-occupier households
%LA Renters	% Local Authority households
%Private Renters	% Private renting households
%No car	% Households with no car
SEG	% Heads of households in SEGs 1 to 5 and 13
%New Comm.	% Persons born in New Commonwealth
%Pensioners	% Persons who are pensioners (men>65, women>60)
%Aged 75+	% Persons aged 75 or over
%Aged 16+	% Persons in private households aged 16+
%Econ. Active	% Persons aged 16+ who are economically active
%Econ. Active Women	% Economically active females (16+)
%Unemployed	% Persons who are unemployed
%Men Unemployed	% Males who are unemployed
%Women Unemployed	% Females who are unemployed
%Car to Work	% Employed persons who travel to work by car
Density	Population density (persons per 10 hectares)

c. SEH and GHS Variables

Twelve key housing variables were identified from the 1993 SEH data set (Table F3). Although three of these variables correspond to equivalent stratifying variables (the proportions of owner-occupiers, private renters and Local Authority renters), they were included in the analyses to show the extent of lost precision should the equivalent census variables not be chosen as stratifiers (in practice Census variables will differ from their respective survey estimates due to differences in their definitions and collection methods, e.g. the stratifier, %LA Renters, also includes housing association renters).

Table F3. Key variables on the SEH with means and design effects

Description	Mean	Deff
% households renting from a housing association	4.1	3.87
% households renting privately	8.4	1.54
% households renting from an LA or New Town	19.9	2.52
% households who are owner-occupiers	67.6	2.07
% households living in terraced housing	28.8	3.29
% households living in flats	17.5	3.06
% households living in detached housing	20.0	2.57
% households living in semi-detached housing	33.2	3.01
% households living in bungalows (1-storey houses)	10.5	2.29
% households living in 2-storey accommodation	74.4	2.65
% households living in accommodation with 3+ floors	15.1	3.66
Number of rooms per person in household	2.6	1.26

The GHS variables were chosen with two intentions; firstly that they covered as wide a subject range as the survey itself does; and secondly that some of the design effects for the variables were relatively large. Broad subject areas covered include health and lifestyles, marital status and family type, education, employment, household definition, and ethnic group (Table F4). It should be noted that some of these subjects, such as health and lifestyles, occur on the survey only every other year.

Table F4. Key variables on the GHS with means and design effects

Description	Mean	Deff
% adults 16-69 educated to at least 'A' level	30.3	1.70
% consulting a GP 2 weeks prior to survey	15.3	1.23
% (adult) frequent light drinkers (QF rating 4)	37.0	1.56
Mean number of people per household	2.4	1.24
% lone parents with dependent child	6.0	1.12
% households with married head of household (HOH)	56.4	1.45
% individuals with limiting long-standing illness	18.7	1.45
% households with non-white HOH	4.1	2.18
Mean length of residence of HOH (yrs)	12.8	1.30
% (adult) current cigarette smokers	27.8	1.56
% adult women who worked in week prior to the survey	49.1	1.35
% adult men who worked in week prior to the survey	63.5	1.13

d. Selecting a regional stratifier for the GHS and SEH

At present the financial surveys (FES & FRS) use a 24-band regional stratifier (FES Region) whilst the GHS stratifies by 22 regional groups. The former differs from the latter only in the way it divides London, dividing the capital into 4 quadrants (NE, NW, SE, SW), whilst the GHS divides it into inner and outer London. The SEH uses the same breakdown as the GHS in England but does not cover any other parts of the United Kingdom, thus having only 15 groups.

For most of the SEH variables, the GHS regional breakdown explained much more of their variation than did FES Region, due to the different partition of London (Table F5). In particular the proportions of owner-occupiers, council tenants, flats, two-storey accommodation and three-storey or more accommodation had particularly high variation between inner and outer London areas, but low variation between the four quadrants. Only for the proportion of private renters did FES Region explain more of the variation than GHS Region, but the difference was small (1%).

In contrast to this, for the majority of the GHS variables, the FES banding explained a greater amount of variance. However, the difference was small in most cases. For the proportion of households with a married HOH, GHS Region was noticeably better at explaining the variance. Overall there was little evidence to warrant choosing one regional stratifier over the other and GHS Region is used here in subsequent analyses.

Table F5: R² for Region on key SEH and GHS variables

	GHS Regions	FES Regions
SEH Variables		
% housing association renters	0.06	0.06
% private renters	0.09	0.10
% LA or New Town renters	0.15	0.12
% owner-occupiers	0.21	0.13
% terraced housing	0.06	0.06
% flats	0.48	0.36
% detached housing	0.25	0.25
% semi-detached housing	0.19	0.15
% bungalows (1-storey houses)	0.19	0.20
% 2-storey accommodation	0.32	0.17
% 3 or more storeys	0.40	0.25
Number of rooms per person	0.07	0.07
GHS Variables		
% A Levels or above	0.07	0.08
% consulted GP	0.05	0.05
% frequent light drinkers	0.16	0.17
Mean household size	0.05	0.05
% lone parents with child	0.08	0.09
% married HOH	0.13	0.10
% limiting long-term illness	0.06	0.06
% non-white HOH	0.25	0.25
Mean length of residence	0.09	0.08
% smokers	0.09	0.08
% working women	0.11	0.12
% working men	0.13	0.14

e. Selecting the second stratifier

In order to select the second stratifier, forward stepwise regression was carried out on a model with GHS Region already entered. The first three variables fitted by this method to each of the key SEH and GHS variables are shown in Tables F6 and F7.

Table F6: Stepwise regression on SEH variables (after GHS Region)

SEH Variables	GHS region +			R ² after 3 steps
	step 1	step 2	step 3	
% housing association renters	%Women Unemployed	%Econ. Active Women	%Aged 16+	0.26
% private renters	%Private Renters	%Aged 16+	%Pensioners	0.48
% LA or New Town renters	%LA Renters	%Women Unemployed	%Car to Work	0.76
% owner-occupiers	%Owners	%Aged 16+	-	0.82
% terraced housing	%No Car	%Pensioners	SEG	0.33
% flats	%Car to Work	%No Car	SEG	0.73
% detached housing	%No Car	%Econ. Active Women	SEG	0.67
% semi-detached housing	%Car to Work	%Women Unemployed	SEG	0.39
% bungalows (1-storey)	%No Car	%Econ. Active Women	SEG	0.35
% 2-storey accommodation	%Car to Work	%Aged 75+	SEG	0.47
% 3 or more storeys	%Car to Work	SEG	%No Car	0.66
Number of rooms per person	SEG	%Pensioners	%LA Renters	0.33

As would be expected, on the SEH each housing tenure variable (private renting, council renting and owner-occupier) fitted their equivalent stratifier first (%Private Renters, %LA Renters, and %Owners). Nearly all the other variables fitted one of two car-related variables, %Car to Work and %No Car. The former was selected first more often than the latter, but since %No Car has been proposed as a stratifier on the FES, this

appeared to be the best overall choice for the first stratification factor. SEG was also chosen on all the non-tenure variables, and was first choice on number of rooms per person.

The choice of best stratifier was much less obvious for the GHS than it was for the SEH, with nine different variables being entered first for the twelve regressions (Table F7). However, SEG was chosen first in three of the regressions and at subsequent steps on a further two. SEG was also highly correlated with FES and FRS variables, and was chosen as the first non-regional stratifier for both these surveys. Therefore, we would not be compromising these other sample designs by selecting it as the first stratifier for the GHS.

Table F7. Stepwise regression on GHS variables (after GHS Region)

GHS variables	GHS region +			R ² after 3 steps
	step 1	step 2	step 3	
% A Levels or above	SEG	%Private Renters	%Econ. Active	0.48
% consulted GP	%Women Unemployed	-	-	0.07
% frequent light drinkers	SEG	%Pensioners	%No Car	0.44
Mean household size	%Aged 16+	%Owners	% New Comm.	0.30
% lone parents with child	%Unemployed	%Aged 16+	%No Car	0.22
% married HOH	%Car to Work	%No Car	%Private Renters	0.39
% limiting long-term illness	%Econ. Active	SEG	%Econ. Active Women	0.17
% non-white HOH	% New Comm.	%Women Unemployed	-	0.77
Mean length of residence	%Private Renters	%Women Unemployed	SEG	0.18
% smokers	SEG	%Owners	%Pensioners	0.37
% working women	%Econ. Active Women	%Women Unemployed	-	0.31
% working men	%Econ. Active	%Women Unemployed	%Aged 75+	0.35

f. Selecting the third stratifier

Having chosen %No Car as the second stratifier for the SEH, stepwise regression was carried out on a model containing GHS Region and %No Car. The results are shown in Table F8. SEG was chosen most consistently across all the non-tenure variables, either as the first choice or the second choice. This appeared to be the best choice as the second stratifier, especially as it is a current stratifier on the FRS and FES. Rather worrying, however, was the absence of both %No Car and SEG in any of the stepwise regressions for the housing tenure variables.

Table F8. Stepwise regression on SEH variables to find the second stratifier

SEH Variables	R ² for GHS Region + %No Car	step 1	step 2	R ² after step 2
% housing association	0.15	%Women Unemployed	%Econ. Active Women	0.25
% private renters	0.10	%Private Renters	%Aged 16+	0.48
% LA or New Town renters	0.41	%LA Renters	%Women Unemployed	0.76
% owner-occupiers	0.61	%Owners	%Aged 16+	0.82
% terraced housing	0.27	%Pensioners	SEG	0.33
% flats	0.66	SEG	%Aged 75+	0.73
% detached housing	0.58	%Econ. Active Women	SEG	0.67
% semi-detached housing	0.31	SEG	%Women Unemployed	0.39
% bungalows (1-storey)	0.28	%Econ. Active Women	SEG	0.35
% 2-storey accommodation	0.35	SEG	%Pensioners	0.47
% 3 or more storeys	0.51	SEG	%Aged 75+	0.66
Number of rooms per person	0.15	%Aged 75+	SEG	0.32

For the GHS, a similar procedure was carried out, but %No Car was replaced by SEG (Table F9).

Table F9. Stepwise regression to choose the second stratifier for the GHS

GHS variables	R ² for GHS Region + SEG	Step 1	Step 2	R ² after step 2
% A Levels or above	0.44	%Private Renters	%Econ. Active	0.48
% consulted GP	0.07	%Women Unemployed	-	0.08
% frequent light drinkers	0.38	%Pensioners	%No Car	0.44
Mean household size	0.05	%Aged 16+	%No Car	0.29
% lone parents with child	0.14	%Women Unemployed	%Aged 16+	0.20
% married HOH	0.16	%No Car	%Private Renters	0.39
% limiting long-term illness	0.11	%Econ. Active	%Econ. Active Women	0.17
% non-white HOH	0.26	%New Comm.	%Women Unemployed	0.77
Mean length of residence	0.09	%Unemployed	%Econ. Active Women	0.18
% smokers	0.30	%Owners	%Pensioners	0.37
% working women	0.16	%Econ. Active Women	%Women Unemployed	0.31
% working men	0.23	%Econ. Active	%Women Unemployed	0.35

A choice of third stratifier was not immediately obvious, so we needed to look at the significance of each variable and to compare the increase in R² which accompanied the addition of each variable to the model containing region and SEG. We limited the analysis to the variables which appear more than twice in the above table, namely %Econ. Active, %Women Unemployed, %No Car and %Econ. Active Women.

The results are given in Table F10 and show that there was little difference in R² between the four fitted variables. The economic activity variables (%Econ. Active and %Econ. Active Women) were significantly correlated with the same seven GHS variables. Between these, %Econ. Active produced the larger values of R². However %Women Unemployed was significantly correlated with nine GHS variables and %No

Car was significantly correlated with all but the two health variables; the proportion of adults who had consulted GP and the proportion with a limiting long-term illness (although these were poor correlates of most of the other census variables too). %No Car was particularly strongly correlated with mean household size and the proportion of households with a married HOH, and substantially increased the R² for the models with these as dependent variables. This was a particularly useful result since SEG had a low correlation with both these variables. So, by choosing %No Car we would not compromise the effectiveness of stratification on the GHS, with the advantage that we will be using a good correlate of variables on the SEH and FES.

Table F10. R² for GHS Region + SEG + third stratifier on key GHS variables

GHS variables	%Econ. Active	%Econ. Active Women	%Women Unemployed	%No Car
% A levels or above	*0.45	*0.44	0.44	*0.45
% consulted GP	0.07	0.07	*0.08	0.07
% frequent light drinkers	*0.40	*0.40	*0.40	*0.40
Mean household size	*0.09	*0.07	0.05	*0.12
% lone parents with child	0.14	0.14	*0.18	*0.17
% married HOH	0.17	0.17	*0.24	*0.37
% limiting long-term illness	*0.16	*0.14	0.11	0.12
% non-white HOH	0.26	0.26	*0.37	*0.30
Mean length of residence	0.10	0.10	*0.16	*0.15
% smokers	*0.32	*0.32	*0.33	*0.32
% working women	*0.27	*0.28	*0.22	*0.20
% working men	*0.32	*0.31	*0.29	*0.27

* significant to 5% level

g. Selecting the fourth stratifier

The fourth stratifier was chosen in a similar way, by comparing the R² for each model containing GHS region, SEG, %No Car and most the remaining variables in turn. Table F11 shows the relative R² for GHS variables for models containing %Men Unemployed, %Econ. Active, %Aged 16+, %Owners, %LA Renters, %Private Renters and %Pensioners⁹.

The results showed that the tenure stratifiers (%Private Renters, %Owners, %LA Renters) were poor correlates of the GHS variables. %Men Unemployed correlated significantly with 8 of the GHS variables, but produced lower values of R² than %Econ. Active, which was correlated significantly with 9 variables. %Aged 16+ and %Pensioners were both correlated significantly with 10 GHS variables, so either one of these or %Econ. Active appeared to be the best general choice.

Table F11. R² for GHS Region + SEG + %No Car + 4th stratifier on key GHS variables

GHS Variables	%Men Unemployed	%Econ. Active	%Aged 16+	%Owners	%LA Renters	%Private Renters	%Pensioners
% A Levels or above	0.45	*0.46	*0.46	0.45	*0.46	*0.48	*0.46
% consulted GP	0.07	0.07	0.07	0.07	0.07	0.07	0.07
% frequent light drinkers	*0.41	*0.43	*0.42	0.40	0.40	*0.40	*0.44
Mean household size	*0.16	*0.14	*0.29	0.12	0.12	*0.13	*0.23
% lone parents with child	*0.18	*0.17	*0.21	0.17	*0.18	*0.17	*0.20
% married HOH	0.37	0.37	*0.38	0.37	0.37	*0.39	0.37
% Limiting long-term illness	0.12	*0.16	*0.14	0.12	0.12	0.12	*0.15
% non-white HOH	*0.32	0.30	*0.36	*0.30	*0.32	*0.32	*0.34
Mean length of residence	*0.17	*0.16	*0.16	0.15	0.14	*0.15	*0.16
% smokers	*0.33	*0.36	*0.34	*0.34	*0.34	0.32	*0.37
% working women	*0.22	*0.28	0.20	0.20	0.20	0.20	*0.24
% working men	*0.28	*0.33	*0.28	0.28	*0.28	0.28	*0.30

* significant to 5% level

⁹ %Pensioners is used in preference to %Aged 75+ and %Econ. Active is used in preference to %Women Econ Active as the preferred variables produced higher R² (result not shown), whilst %Men Unemployed is used in preference to %Unemployed and %Women Unemployed since this is a proposed stratifier on the FES.

For the SEH variables, models containing the housing tenure variables not surprisingly produced some of the highest values of R², but these were for models containing the corresponding survey tenure variables as dependent variable (Table F12). For the more general housing variables, the values of R² were not as high as were produced by the other models. %Men Unemployed was again a poor correlate of many of the SEH variables, whilst %Econ. Active and %Aged 16+ produced generally high values of R² for most of the variables. However, %Pensioners accounted for the highest or joint highest amount of variance in over half of the non-tenure SEH variables. Interestingly %Pensioners correlated more highly with the survey variable for the proportions of private renters, than did %Owners, and with the survey variable for the proportions of owner-occupiers, than did %Private Renters. Therefore %Pensioners would appear to be the best overall choice as last stratifier, especially since it was also one of the best choices in terms of the GHS variables. However, there is an argument for using either %Private Renters or %LA Renters as a fifth stratifier to increase precision on tenure variables.

Table F12. R² for GHS Region + SEG + %No Car + 4th stratifier on key SEH variables

SEH Variables	%Men Unemployed	%Econ. Active	%Aged 16+	%Owners	%LA Renters	%Private Renters	%Pensioners
% housing association	*0.22	*0.19	0.18	*0.19	*0.18	0.18	*0.19
% private renters	0.15	0.14	*0.15	0.14	*0.24	*0.41	*0.15
% LA or New Town renters	0.46	0.46	*0.47	*0.64	*0.76	*0.60	*0.47
% owner-occupiers	*0.62	*0.62	*0.63	*0.82	*0.76	*0.62	*0.63
% terraced	*0.30	*0.32	*0.30	*0.30	*0.31	*0.30	*0.33
% flats	0.72	0.72	*0.72	*0.72	0.72	*0.73	*0.73
% detached	*0.64	*0.66	*0.66	*0.61	*0.62	*0.61	*0.61
% semi-detached	*0.38	0.37	0.37	*0.38	0.37	*0.39	0.37
% bungalows	*0.33	*0.35	0.32	*0.32	*0.32	0.32	*0.33
% 2-storey	0.42	*0.44	*0.44	*0.43	0.42	*0.44	*0.47
% 3+ storeys	0.64	*0.64	*0.65	*0.64	*0.64	*0.66	*0.66
Number of rooms per person	*0.21	*0.29	*0.25	*0.21	*0.22	*0.21	*0.33

* significant to 5% level

To check the suitability of the stratifiers for income and expenditure variables (from the FES, although income variables will be relevant to the FRS) we repeated the analyses for 11 income and expenditure variables, using GHS region for comparability (Table F13). For all of the expenditure variables and most of the income variables, there was a poor correlation with the majority of the census variables tested. There was little fluctuation in R^2 across all the models. For benefit income, %Men Unemployed and %Econ. Active were better than the others in explaining variation, and the latter also explained a greater amount of variation in income from pensions than did %Pensioners (which nevertheless still increased R^2 by a significant amount). %Men Unemployed was significantly correlated with just three variables. This was recommended (but was not used) as the fourth stratifier for the FES because of its strong correlation with gross income, and only one other tested variable is a significant predictor of this (%Aged 16+). %Econ. Active was the strongest predictor of other income variables (these variables were not used in the previous analysis for the FES) and was also the best predictor of motoring expenditure. Few of the other tested variables were correlated with any of the expenditure variables.

Table F13. R^2 for GHS Region + SEG + %No Car + 4th stratifier on key FES variables

FES Variables (variables are means for PSUs)	%Men Unemploy ed	%Econ. Active	%Aged 16+	%Owner s	%LA Renters	%Private Renters	%Pension ers
gross income	*0.44	0.43	*0.44	0.43	0.43	0.43	0.43
earnings (wages)	0.36	*0.37	0.35	*0.36	0.36	0.35	*0.36
self-employ income	*0.08	*0.08	*0.09	0.07	0.07	*0.09	*0.08
benefit income	*0.23	*0.23	*0.20	*0.20	*0.21	*0.21	0.20
pensions income	0.17	*0.26	*0.19	0.17	0.18	0.17	*0.22
total expenditure	0.43	0.43	0.43	0.43	0.43	0.43	0.43
food expenditure	0.30	0.30	0.30	0.30	0.30	0.30	0.30
motoring expenditure	0.17	*0.18	0.17	0.17	0.17	*0.17	0.17
housing expenditure	0.42	0.43	0.42	0.42	0.42	0.42	0.43
h/h goods expenditure	0.15	0.15	0.15	0.14	0.15	0.15	*0.15

Over all surveys the choice seemed to be between %Econ. Active, %Aged 16+ and %Pensioners. %Econ. Active correlated well with variables related to income and expenditure, economic activity, and non-tenure housing variables, whilst %Aged 16+ and %Pensioners were both highly correlated with a wide range of variables. Only on a few variables did the choice of fourth stratifier make a large difference to R^2 , so unless there was a case for choosing a variable which was particularly strongly correlated with

a specific survey variable, it would be best to choose a good overall variable such as %Pensioners as fourth stratifier.

h. Comparison of proposed stratifiers with current stratifiers

We compared models containing the proposed IHS stratifiers (with %Pensioners as fourth stratifier) with the models containing the current stratification factors from the SEH/GHS, the FRS and the FES, on key survey variables from the SEH, GHS and FES.

The proposed IHS stratifiers were as good or better than the current GHS stratifiers on all but one (the proportion of adults with A-levels or above) of the GHS variables (Table F14). The biggest improvement was for mean household size, which did not seem to be well correlated with either of the census tenure variables. This increased R^2 by 12 percentage points. Additionally, only for the proportion of non-white HOHs was the R^2 for the optimum stratifier model noticeably higher than that for the proposed model. This was due to its correlation with % New Comm. On this evidence it seemed that there would be little or no loss to precision of GHS estimates should the proposed IHS stratifiers be adopted.

Table F14. Comparison of R^2 for current GHS stratifiers and proposed IHS stratifiers on key GHS variables

GHS Variables	Optimum R^2 for 4 stratifier model (from Table 6)	Current GHS stratifiers (GHS Region + %Private Renters + %LA Renters + SEG)	Proposed IHS stratifiers (GHS Region + SEG + %No Car + %Pensioners)
% A Levels or above	0.48	0.48	0.46
% consulted GP	0.07	0.07	0.07
% frequent light drinkers	0.44	0.40	0.44
Mean household size	0.30	0.11	0.23
% lone parents with child	0.22	0.17	0.20
% married HOH	0.39	0.35	0.37
% Limiting long-term illness	0.17	0.12	0.15
% non-white HOH	0.77	0.30	0.34
Mean length of residence	0.18	0.16	0.16
% smokers	0.37	0.35	0.37
% working women	0.31	0.19	0.24
% working men	0.35	0.26	0.30

For the SEH variables, the proposed IHS stratifiers did, in general, improve R^2 over that obtained by the model for current SEH variables (Table F15). Only for three of the four tenure variables (those with equivalent census variables - the proportions of private renters, LA renters and owner occupiers) did the value for R^2 fall markedly short of this.

Surprisingly, on five of the SEH variables, R^2 for the model containing the proposed IHS stratifiers was the same as that for the model containing the optimum four stratifiers. On the other variables (except the three tenure variables), there was a drop in the amount of variance explained, but the differences between the two models were not great. We can conclude, therefore, that the proposed stratifiers would not reduce precision except in a few specific cases.

Table F15. Comparison of R^2 for current SEH stratifiers and proposed IHS stratifiers on key SEH variables

SEH Variables	Optimum R^2 for 4 stratifier model (from Table F6)	Current SEH stratifiers (GHS Region + %Private Renters + %LA Renters + SEG)	Proposed IHS stratifiers (GHS Region + SEG + %No Car + %Pensioners)
% housing association renters	0.26	0.19	0.19
% private renters	0.48	0.41	0.15
% LA or New Town enters	0.76	0.75	0.47
% owner-occupiers	0.82	0.81	0.63
% terraced	0.33	0.28	0.33
% flats	0.73	0.69	0.73
% detached	0.67	0.57	0.61
% semi-detached	0.39	0.37	0.37
% bungalows	0.35	0.26	0.33
% 2-storey	0.47	0.45	0.47
% 3+ storeys	0.66	0.63	0.66
Number rooms per person	0.33	0.22	0.33

Differences in R^2 between models containing the FES stratifiers and the proposed IHS stratifiers were very small (Table F16), since the stratifiers themselves were similar. The main difference occurred for pension income due to the effect of %Pensioners. As the FRS stratifiers differed more from the IHS stratifiers, so did the amount of variance explained. However the differences were again slight, with the major differences being for benefit income and pension income again.

Table F16. Comparison of R^2 for current FRS and FES stratifiers and proposed IHS stratifiers on selected FES variables

FES Variables	Current FRS stratifiers (FES Region + SEG + %Unemployed + %Owners)	Current FES stratifiers (FES Region + SEG + %No Car)	Proposed stratifiers (GHS Region + SEG + %No Car + %Pensioners)
gross income	0.43	0.43	0.43
earnings (wages)	0.38	0.37	0.36
self-employed income	0.08	0.07	0.08
benefit income	0.23	0.19	0.20
pensions income	0.17	0.18	0.22
total expenditure	0.43	0.43	0.43
food expenditure	0.29	0.30	0.30
motoring expenditure	0.17	0.17	0.17
housing expenditure	0.42	0.42	0.43
household goods expenditure	0.13	0.14	0.15

Publications in the GSS Methodology Series

Recent publications

(price £5 unless otherwise stated)

- Report 1 Dave Elliot
Software to weight and gross survey data with applications to the EC Household Panel and Family Expenditure Surveys
- Report 2 *Report of the task force on seasonal adjustment*
- Report 3 *Report of the task force on disclosure*
- Report 4 *Report of the task force on imputation*
- Report 5 Peter Sharp
GDP: Output methodological guide
price £20
- Report 6 Michael A. Baxter
Interpolating annual data into monthly or quarterly data

Forthcoming publications

- Andrew Allan
Report on the improved overseas trade in services survey methodology
- Kate Foster
Evaluating non-response in household surveys.

Available only from the Office for National Statistics - see page ii for contact points.
Copies available free of charge to GSS members.