Understanding Society
Working Paper Series

No. 2009 – 02

December 2009

Design of the *Understanding Society* Ethnic Minority Boost Sample

Richard Berthoud, Laura Fumagalli, Peter Lynn, Lucinda Platt

Institute for Social and Economic Research
University of Essex

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

Understanding Society
THE UK HOUSEHOLD LONGITUDINAL STUDY

Design of the *Understanding Society* Ethnic Minority Boost Sample

Richard Berthoud
Laura Fumagalli
Peter Lynn
Lucinda Platt
Institute for Social and Economic Research
University of Essex

Abstract

The overall sample design for *Understanding Society* has been described in an earlier working paper in this series (Lynn 2009). This paper describes the special measures taken to boost the sample of members of five key minority ethnic groups in Great Britain. A new method was developed to estimate the ethnic density of postal sectors in 2007, when the most recent Census data was collected in 2001. Key stages from then on were: excluding sectors with low minority populations; selecting addresses using fractions which over-sampled areas with high densities of the scarcest groups; sub-sampling sectors with low expected yields; and screening in the field to identify target households.

Key words: sample design, panel surveys, ethnic minorities

JEL classification: C83, J15

Author contact details: berthoud@essex.ac.uk

Acknowledgements: The work reported in this paper is part of the “Understanding Society” programme, funded by the ESRC. Implementation of the sample design was carried out by the National Centre for Social Research (NatCen).

Thanks to Susan Purdon, then of NatCen, whose initial proposal for the design of the LSEM (Nazroo and others 2005) was the starting point for the actual design of the UKHLS boost sample described here; and to Kevin Pickering and Sarah Tipping of NatCen for important contributions to the detail of the sample selection procedure. Thanks to Vivienne Avery, Paul Chapman and Laura Keyse of the Office for National Statistics for making available data from the Annual Population Survey. Thanks to UK Geographics for supplying the Census and PAF data about postal sectors.

Design of the *Understanding Society* Ethnic Minority Boost Sample

Non-technical summary

A key objective of *Understanding Society* was to provide detailed comparisons of social and economic experiences across ethnic groups, as well as to study issues of special relevance to ethnic minorities. Although the general population sample and the BHPS included substantial numbers of members of ethnic minority groups, it was decided to boost the sample. The specific objective was to add at least 1,000 adults from each of five communities: Indians, Pakistanis, Bangladeshis, Caribbeans and Africans.

The boost sample was designed around variations in ethnic densities in small areas within Great Britain. A first requirement was to obtain the best estimate of the density of each of the target groups within every small area. These estimates were based on a new technique cross-analysing 2001 Census data for postal sectors with micro-data from the Annual Population Survey.

The aim was to boost the number of addresses selected in areas of high concentration, and to curtail the number selected in areas of low concentration. The three main processes were:

- Excluding areas of very low minority density from the sample altogether
- Varying the sampling fraction within small areas so that a large proportion of addresses would be selected in areas of high density (especially of the scarcest target groups).
- Sub-sampling of areas in which a very small number of interviews with members of target groups was predicted.

Once the addresses had been selected, interviewers were asked to visit each address to find out whether any members of the targeted or included minority groups lived there. This is known as screening. Where relevant minority groups were identified, there was a secondary selection process whereby all households containing members of the scarcest minority groups were recruited to the survey; but some households containing members of the most common minority groups were deselected at random to reduce their sample size to the target number.

Weighting will be required at the analysis stage to counter-balance the variations in selection probabilities between postal sectors and between ethnic groups.

Contents

Introduction

Structure of the survey as a whole

Understanding Society is the major new panel survey of British households. It is funded by the Economic and Social Research Council, with substantial support from the Department for Business, Innovation and Skills. The study is based at, and led by, the Institute for Social and Economic Research (ISER) at the University of Essex, working with colleagues from the University of Warwick and the Institute of Education. The survey fieldwork is being delivered by the National Centre for Social Research (NatCen). *Understanding Society* both replaces and incorporates the much smaller British Household Panel Survey (BHPS), which has been running since 1991.

Key characteristics of the new survey include:

- a total target sample size of 40,000 households, with four main components:
    - an innovation panel
    - a new national equal probability sample
    - an ethnic minority boost sample
    - incorporation of the existing sample from the BHPS.
- interviews with all household members, aged 10 and above
- topic coverage relevant to a wide range of disciplines and policy fields
- an ethnicity strand, focussing on the identity and social position of minority groups
- collection of health indicators and biomarkers
- links to supplementary data, such as neighbourhood information
- a platform for the collection of qualitative data
- an Innovation Panel for methodological research.

Understanding Society is a 'household panel survey'. A large representative sample of households has been selected across the United Kingdom. The households are being visited by an interviewer in a rolling first wave starting in January 2009, and all the adult household members are asked detailed questions about a range of subjects: family structure, employment, income, health and so on. Each member of the sample is then re-interviewed a year later, to see how things have changed over the past 12 months; and again and again in subsequent years for as long as the survey lasts. This 'longitudinal' approach provides much clearer evidence about the processes underlying social and economic change, and enables analysts to make inferences about causation which cannot be supported by one-off, cross-sectional surveys.

This paper describes the design of the ethnic minority boost sample, a major element of the ethnicity strand of the survey.

Historical background to the ethnicity strand

The historical background to the new survey as a whole runs from the establishment of the Panel Study of Income Dynamics (PSID) in the US in the early 1970s (psidonline.isr.umich.edu), through the German Socio-Economic Panel (GSOEP) in the 1980s (www.diw.de/english/soep) and the British Household Panel Survey in the 1990s (www.iser.essex.ac.uk/survey/bhps)

The historical background to the ethnicity strand has run a different course, over roughly the same period. A series of detailed surveys of Britain's ethnic minorities was carried out at roughly ten year intervals by the Policy Studies Institute. The first focussed narrowly on their experience of discrimination (Daniel 1967); the second expanded the enquiry to examine the wider concept of disadvantage (Smith 1976); the third compared ethnic minorities as a group with the white population across a wide range of issues (Brown 1984), while the fourth took a more pluralistic view, comparing the distinct experiences of each minority group (Modood, Berthoud and others 1997). Each successive survey widened the range of topics investigated, and of minority groups covered.

At the turn of the century, consideration was given to the possibility of undertaking a fifth national survey of ethnic minorities. But by this time several large-scale national surveys supported by government departments enabled analysts to make detailed comparisons between ethnic groups, either because of the very large general sample sizes accumulated over a series of years (for example Labour Force Survey, the Family Resources Survey) or because of targeted boost samples of minority groups (the Health Survey for England (in 1999 and 2004), the British Crime Survey (in 1988, 1992, 1996 and 2000), the Citizenship Survey (since 2001).[1] It was felt that these new surveys largely met the current need for cross-sectional data about the main ethnic minority groups. But a review of longitudinal data resources carried out by ONS in 2000 concluded that the BHPS had too small a sample for serious analysis of ethnicity. The ONS Longitudinal Study linking the 1971, 1981, 1991 and 2001 Censuses provided adequate sample sizes, but was restricted both in the subject matter covered and in the long gap (10 years) between available observations. There was an urgent need for a longitudinal survey of ethnic minorities (LSEM).

ONS and the ESRC commissioned first a scoping study (Owen and Green 2003) and then a feasibility study (Nazroo and others 2005) to examine the options for a dedicated longitudinal survey of ethnic minorities, with, probably, a longitudinal survey of the majority white population to provide a comparison group. This plan was initially adopted by the ESRC, and a substantial budget was set aside to fund it.

The plan for a stand-alone panel survey of ethnic minorities was then incorporated in the even more ambitious proposal for a UK household longitudinal survey, later to be known as *Understanding Society*. Within the overall objectives of the new survey, covering a very wide range of question topics, the 'ethnicity strand' has been designed to enable analysis of ethnicity and comparison of individual ethnic minority groups through the following means:

1. a substantial number of members of minority ethnic groups within the main equal probability sample;
2. a further substantial boost sample of members of five key minority groups;
3. the opportunity for analysts to focus on variations between ethnic groups in outcomes of universal interest such as employment, income, housing, health and so on.

[1] Details of these surveys are available at http://www.data-archive.ac.uk/findingData/majorstudies.asp

4. inclusion of some questions of specific interest to the study of ethnicity, such as ethnic identity, religious and cultural attitudes, the experience of discrimination and harassment, and so on.

This working paper outlines the methods used to select the boost sample – item 2

Overall design of the boost sample

Understanding Society is a household panel survey – that is, a sample of households is selected in the base year, data is collected from or about all members of each household, and each member is followed up in subsequent years.

The main sample of 28,000 households provides the benchmark data for the whole population, and for the white majority, against which minority ethnic experiences should be compared. Unlike previous specialist surveys of the ethnic minorities, there is no need for a 'white comparison sample', because that is built into the overall design.

The main sample also includes an estimated 2,800 adults from the groups covered by the boost sample, plus some others from groups not covered by the boost (for example white minorities). Although there would be a five-minute sequence of specialist questions asked only of the boost sample, the principal aim was to combine the boost sample with the main sample to provide large numbers of members of minority groups for analysis of the full range of survey data.

The twin objectives of the boost sample design were:

- to select an additional sample of households containing members of five target ethnic minority groups (Indians, Pakistanis, Bangladeshis, Caribbeans and Africans) plus such other minority groups as the sampling procedure allowed;
- to arrange the selection in such a way that 1,000 to 1,125 adults in each of the target groups would be interviewed in the first wave.

See a later section of this paper (page 5) for detailed definitions of the ethnic groups to be covered.

The first, general, objective has been addressed in earlier surveys referred to on the previous page. The second, group-specific, objective was largely new. The design was broadly based on the recommendations of the feasibility study for an LSEM (Nazroo and others 2005)[2], although there were several differences of detail between the original proposal and the design actually adopted.

The primary design strategy was to focus the sample selection on areas (postal sectors) where members of the target minority groups are known to live in high concentrations.

[2] The sample design section of the LSEM feasibility study was drafted by Susan Purdon, then of the National Centre for Social Research. Her particular contribution was the variation in the sampling fractions between sectors with different ethnic compositions, described as Step 3 below.

If all members of the target groups lived in areas where they were the only residents, locating a sample of them would be a simple matter. Such high levels of ethnic segregation – labelled ‘ghettoes’ - were approximated in American cities such as Chicago in the 1940s (Duncan and Duncan 1950), but have never been replicated in Britain (Peach 1996). Most members of minority ethnic groups live in areas where most other residents are white. Nevertheless, it has long been established that there is sufficient variation in ethnic densities at the small area level to provide leverage for an efficient sample design.

The basic ingredients of such a design are two: a) identification of small areas with high minority concentrations, and b) screening at the fieldwork stage to identify and interview members of the target groups. Neither of these ingredients is efficient on its own.

- Screening undertaken equally across all small areas (rather than concentrated in high density areas) would yield an average of one minority household in every 20 addresses issued, including many fieldwork assignments where the yield would be as low as one in 100. So it was important a) to limit the sample to areas of above average concentration, and b) to focus the sample on areas with very high concentrations, especially of the hard-to-reach target groups.
- Drawing a sample of households in high density areas (without screening) would yield a sample which either included a very large proportion of white households (not of interest to the boost sample) or focussed on a very small and unrepresentative group of minority households who lived in areas of exceptionally high density.[3]

These remain the two essential ingredients, even though the task on this occasion was to select samples of five specific minority groups, rather than simply a single sample of minority groups taken as a whole.

Previous surveys of ethnic minorities (for example Brown 1984, Modood, Berthoud and others 1997) have used the technique known as ‘focussed enumeration’, in which the residents of a sample of addresses were asked whether members of any ethnic minority lived in the neighbouring houses or flats (Brown and Ritchie 1982). It was decided not to use this method on this occasion, even in areas of very low density, in view of recent evidence that the proportion of neighbours who were reported to be members of ethnic minorities was substantially lower than the proportion directly observed at the sample addresses (Smith and others 2010).

[3] Although the requirement to undertake screening as well as oversampling of areas of high density is well established, it is not always followed. The Millennium Cohort Survey, for example, tried to use area variations in sampling fractions to boost the number of ethnic minority (and low income) children in its sample, without screening. It can be shown that once weighting has been applied, this approach actually yields *effective* sample sizes of minority ethnic groups (and of low income families) that are smaller than would have been achieved if the same total number of interviews had been spread over an equal probability sample.

Relationships between components of the overall sample, in contributing to the ethnicity strand

The *Understanding Society* sample is made up of three main components[4]:

- A new general population sample (GPS) expected to consist of about 28,000 households (see Lynn 2009 for details)
- Continuation of the former BHPS sample beyond wave 18, now consisting of about 6,400 households
- The ethnic minority boost sample expected to consist of about 4,200 minority households

Although this paper focuses on the boost, all three components of the sample contribute to the ethnicity strand.

First, the general population and BHPS samples provide data about the majority white population for comparison with minority groups on the main question sequences.

Second, the general population and BHPS samples include minority ethnic groups in their due proportion (expected to total 4,800 adults), and these can be combined with the boost sample (6,800 adults) to enhance the coverage of minorities. Weighting factors will need to be calculated before the three components can be combined for analysis.

Most of the questionnaire addressed to the three components of the sample is in common. But an additional short sequence of questions has been designed specifically for the ethnicity strand. This sequence, known as the "extra five minutes", is asked of all members of the boost sample.

The *third* crossover between the samples is that members of ethnic minorities identified in the general population sample in areas of low minority density (below 5 per cent) are asked the "extra five minutes". This is because areas of low density have been excluded from the boost sample (see page 13), and the GPS provides coverage of sparsely-grouped minorities to enable construction of a sample genuinely representative of all minorities.

Fourth, a small sub-sample of about 600 responding households in the GPS is asked the "extra five minutes" (even if consisting entirely of white people), to provide a comparison between the minorities and the general population on this particular question sequence. This general population comparison sample was chosen by selecting one address from the list of addresses selected at each of 1,056 sampling points. (See Lynn 2009, p. 5)

Defining target ethnic groups

Most previous surveys with a sole or special interest in ethnicity have selected a sample of the 'visible' minority groups taken as a whole, without differentiation

[4] A fourth component, the Innovation Panel, will not be combined with the other three, and makes little direct contribution to the ethnicity strand

between groups of diverse ethnic origin in the sample design (although it has always been possible to make analytical comparisons between groups if the sample sizes were large enough). The fourth national survey (Modood, Berthoud and others 1997) was a partial exception in that special measures were taken to maximise the number of Bangladeshis in the sample – the smallest of the identified target minority groups.

One of the central analytical conclusions of the fourth national survey was that the main individual ethnic minority groups experienced diverse outcomes, which could not be summarised by combining all together under the single label 'minority'. The feasibility study for the LSEM (Nazroo and others 2005) confirmed this emphasis on diversity, recommending equal-size samples of each of five target minority groups, rather than a representative sample across all groups - which would have yielded 'too many' members of the largest group (Indians) and 'too few' of the smallest group (Bangladeshis).[5]So the initial specification for the boost sample required that at least 1,000 adults should be interviewed in each of the following ethnic groups, in addition to those located in the main equal probability sample:

- Indians
- Pakistanis
- Bangladeshis
- Caribbeans
- Africans

In explaining the design of the boost sample, it is important to distinguish clearly between three distinct (but obviously related) ways of defining ethnic groups.

The ONS definition: The standard ONS classification of 14 ethnic groups has been used in the Census, the Labour Force Survey, the Annual Population Survey and so on. (They are listed in Table 1 below, though not in the conventional order.) As explained below, the design of the boost sample was based on calculations from Census data supplemented by APS data. So the selection of sampling points, and the estimated number of interviews in each target group, were necessarily based on this definition.

The screening definition: We did not think that the Census classification provided an ideal way of defining ethnic groups – partly because the preamble was unclear about how people should decide what an 'ethnic group' consisted of, and partly because the 14 categories were often too complex for our intention to focus on five target groups. So a different set of questions (ie not the ONS question) was used in the field to define the target groups, and so enable interviewers to decide which households should be included in the boost sample. This was intended to relate to the ONS definition, but both the introductory question and the response categories were designed specially for the screening operation. The screening question was as

[5] The increasing emphasis on diversity could go on and on. The group that was once labelled South Asians is now routinely split into Indians, Pakistanis and Bangladeshis, with huge differences in observed outcomes. But it can be argued that 'Indian' is a geo-political category that should ideally be split into sub-groups defined by region of origin, religion or language. Similarly, although 'Africans' have been introduced as a target category for the first time in the current sample design, we can expect wide diversity in the UK experiences of people migrating from, for example, Nigeria or Somalia.

follows. A fuller version is provided in the screening questionnaire reproduced in Appendix 1.

> Does anyone living at this address come from, or have parents or grandparents from any of the following ethnic groups?
>
> Indian
>
> Mixed Indian[6]
>
> Pakistani
>
> Bangladeshi
>
> Sri Lankan
>
> Caribbean / West Indian
>
> Mixed Caribbean/West Indian[3]
>
> North African
>
> Black African
>
> African Asian
>
> Chinese
>
> Far Eastern[7]
>
> Turkish
>
> Middle Eastern / Iranian[8]
>
> None of these

Analysis definitions: While the ONS and screening definitions between them determined the structure of the sample, analysts are free to adopt a range of possible definitions of each group based on the various questions about country of origin, subjective identity, religious affiliation and so on. Note that all these questions are asked of the whole sample (not just the ethnic minority boost) so that alternative analysis definitions can be tested and applied across all members of minority groups taking part in the survey. The ONS and screening definitions are among the options available for this purpose, but do not constrain the analytical approach.

Bangladeshis are the smallest of these five ethnic minority groups, and it was decided to design the sample in such a way as to aim to achieve an estimated 1,000 interviews with adult members of that community, but at least 1,100 interviews with members of the four larger groups. Among Indians and Caribbeans, there are sizeable sub-

[6] Mixed Indian was defined in the questionnaire as parents or grandparents from Indian ethnic group *and* parents or grandparents from a non-Indian ethnic group. Mixed Caribbeans were defined equivalently.

[7] Examples given were Filipino, Thai, Malaysian, Japanese, Vietnamese, Singaporean, Indonesian, Korean, Burmese

[8] Examples given were Israeli, Palestinian, Lebanese, Syrian, Jordanian, Yemeni, Saudi, Iraqi, Afghani, other Gulf states

minorities with mixed parentage (usually mixed white-Indian and mixed white-Caribbean). The sample design was adjusted to maximise the number of members of these mixed sub-groups, and the target was raised to 1,125 to allow for this.[9]

Among people who were born in Africa (or whose parents had been born in Africa) a sizeable minority have parents or grandparents originating in the Indian sub-continent. This group (referred to in previous surveys as 'African Asians') have been included in the boost sample of Indians (or, as appropriate, of Pakistanis or Bangladeshis) rather than treated as (black) Africans.

Analysis of Census data on country of birth by ethnic group had shown that people born in North Africa are recorded by the Census mostly as either 'white' or 'other'. It was decided to add North African to the list of ethnic groups in the screening questionnaire, and combine them with (black) Africans in the boost sample.

People born in the middle east (including Turkey and Iran as well as 'Arab' countries) also tend to split between 'white' and 'other' when nominating an ethnic group in the Census. This consideration led to inclusion of the following sub-categories in addition to the target groups: Chinese, other far eastern, Sri Lankan, Turks, other middle eastern. Each would yield a few hundred cases in the boost sample, with 1,100 allowed for these various other included groups in total.

Small numbers of non-white minorities with diverse origins were not covered at all in the boost sample – these could include, for example, Australian Aborigines, Maoris, Pacific Islanders, native N and S Americans.

The precursor studies and the ESRC brief had defined the target groups as *non-white* minorities. The research team and its advisors reviewed the considerations for and against adding 'white' minorities to the boost sample. Options discussed ranged from a narrowly defined east-European group, a more broadly defined all-European group, or covering all white minorities including those with recent origins in North America, Australasia or Southern Africa. ('White' people originating in North Africa and the Middle East - mainly Turks and Arabs - have been discussed above.) White minorities are substantial, increasing, and potentially interesting groups. On the other hand, the dividing line between white people with UK and other origins is not easy to establish rigorously; the dividing line may disappear at the second or third generation following migration; many recent migrants are thought likely to return home (and so leave the panel); the technique of locating the boost sample in areas of high minority concentration would not have worked for white minorities; and adding them to the boost would have increased costs (unless the boost was reduced in some other dimension). While recognising the potential desirability of analysing white minorities, it was decided to not to include them in the boost. Of course members of these groups appear in the main sample.

[9] Although the layout of the Census questionnaire, and many of its output tables, encourage analysts to think of 'mixed' as a single ethnic group with common experiences, we hypothesised that people of mixed white/minority heritage might have more in common with the specific minority ethnic group from which one of their parents originated, than with mixed-heritage people from different backgrounds. So we conceive of (for example) mixed white/Caribbean as linked to the Caribbean group, rather than to mixed white/Indian . Analysis will allow this hypothesis to be tested.

If white minorities originating outside the UK were not boosted, it followed that Gypsies/Roma/travellers should not be boosted either. The numbers were likely to be small in any case, and the arrangements for locating them problematic. It was felt that a small dedicated survey of Gypsies/Roma/travellers would be more effective at identifying the particular needs and circumstances of these groups.

The full framework of minorities targeted by the design is shown in Table 1.

Table 1 Targeted, included and excluded groups for the ethnic minority boost sample

Main category	**Census categories used for estimating densities**	**Screening categories used for selection**	**Target**
Target groups			
Indian	Indian Mixed white/Asian	Indian Mixed Indian African Asian	1125
Pakistani	Pakistani	Pakistani	1100
Bangladeshi	Bangladeshi	Bangladeshi	1000
Caribbean	Caribbean Mixed white/Caribbean Black other	Caribbean/West Indian Mixed Caribbean/ West Indian	1125
African	Black African Mixed white/black African	North African Black African	1100
Included groups			
Other included groups	Chinese Other Asian	Sri Lankan Chinese Other far eastern Turkish Middle eastern/Iranian	(1100)
Excluded groups			
Other non-white minorities with diverse origins	Other	None	0
White minorities	White other	None	0

Steps in the design of the boost sample

Overview

As discussed, the *Understanding Society* boost sample was designed around variations in ethnic densities in small areas within Great Britain. The main new requirement was to take account of variations in the location of each of the five specific target groups, rather than targeting all minorities combined.

A first requirement was to obtain the best estimate of the density of each of the target groups within every small area (Step 1 below)

The objective was to boost the number of addresses selected in areas of high concentration, and to curtail the number selected in areas of low concentration. But other considerations apply to limit the extent of this targeting. (At the extreme it would not have been legitimate simply to interview 1,000 Bangladeshis in the most densely concentrated area of Tower Hamlets.) The three main processes were:

- Excluding areas of very low minority density from the sample altogether (Step 2 below)
- Varying the sampling fraction within small areas so that a large proportion of addresses would be selected in areas of high density (especially of the scarcest target groups). This process is described as Step 3 below, and the implications for the efficiency of the design assessed at Step 8.
- Sub-sampling of areas in which a very small number of interviews with members of target groups was predicted (Step 4 below).

Once the addresses had been selected, interviewers were asked to visit each address to find out whether any members of the targeted or included minority groups lived there. This is known as screening (Step 6 below). Where relevant minority groups were identified, there was a secondary selection process whereby all households containing members of the scarcest minority groups were recruited to the survey; but some households containing members of the most common minority groups were deselected at random to reduce their sample size to the target number.

In detail, the following steps were required

Step 1: Estimate the ethnic composition of small areas
Step 2: Exclude postal sectors with very low minority densities
Step 3: Estimate the number of addresses to be selected, using a sampling fraction weighted towards sectors containing the scarcest target groups.
Step 4: Cluster low-yield sectors
Step 5: Select addresses
Step 6: Screening
Step 7: Following rules
Step 8: Weighting

These eight steps will be discussed in turn in the following sections

Step 1: Estimate the ethnic composition of small areas

It was decided to use the Postal Address File (PAF) as the sampling frame of addresses, and this led to the choice of postal units for the geographical analysis of small areas. The sample design was based on postal sectors.[10] There are about 9,000 of these in Great Britain, containing an average of 2,500 households each. Postal sectors are similar in size to electoral wards, which were commonly used as the basis for sample design before the PAF was adopted as the usual sampling frame.

The 2001 Census provides data about the ethnic composition of all postal sectors, analysed according to the standard 14-category classification. The classification was collapsed to seven minority categories for the purpose of sample design calculations as shown in Table 1.

In practice a direct estimate of the number of ‘white’ people in each sector was never needed, as estimates of the size of the minority groups were always expressed as a proportion of the total population.

[10] Postal sectors are defined by the postcode, omitting the final pair of alphabetical characters, eg CO4 3

The census estimates for Africans and the other included groups were adjusted on the basis of an analysis of Census data cross-analysing ethnic group and country of birth.

- The 'African' group was assumed to be 11 per cent larger than the Census black African and mixed white-black African groups, to allow for the fact that white and 'other' people with origins in north Africa would count in the sample as Africans
- The other included groups (taken as a whole) were assumed to be 8 per cent larger than the Census combined 'other mixed', 'other Asian' and 'other' groups to allow a) for the fact that white people with origins in the middle east would be included in the sample, while on the other hand b) some 'other' small minorities with diverse origins would not contribute to the sample at all.

These adjustments were designed to improve the accuracy of the targeting, but make no difference to the 'screening' or 'analysis' definitions of ethnic groups referred to on pages 6 and 7.

Census data for the number of adults in each postal sector in Great Britain provided the main base for sample selection. Small adjustments had to be made to the figures for Scotland to take account of the slightly different ethnic classification in the Scottish Census. Northern Ireland was not covered by the boost sample, as so few members of the target minority groups live there.

We were concerned that data collected in the 2001 Census might not provide an accurate estimate of the distribution of minority ethnic groups at the time the sample was selected in 2008. This is a cyclical problem for sample designs based on Census data, especially acute late in the 10-year sequence between Censuses. In order to improve the estimates, we asked ONS to provide us with the raw individual-level data showing ethnic group and postal sector from the Annual Population Survey (APS) for 2007, under special license.

The APS covered about 150,000 households. Since this amounts to only about 16 households per postal sector, on average, it does not provide accurate estimates of the ethnic distribution of the population of specific sectors. It was used, though, to provide estimates of the rate of change in ethnic composition, using the 2007 APS distribution as the dependent variable, and the 2001 Census distribution as the predictor variable. This enabled us to estimate the 2007 distribution for each postal sector, uprating the 2001 figures by factors derived from the 2007 analysis.

For each of the five target groups, a series of regression models was run in which each sector's 2007 proportion of adults in the relevant ethnic group was predicted on the basis of:

- the 2001 proportion in that ethnic group
- the square of the 2001 proportion
- a set of conurbation dummies: inner London, outer London, est Midlands, West Yorkshire, Greater Manchester (all other regions combined)
- interactions between conurbation and the 2001 proportion .

The use of the term for the square of the 2001 proportion was designed to test for the possibility that minority densities tended either to rise fastest in areas of already high density (increasing concentration) or in areas of previously low density (increasing dispersal).

Table 2 shows the coefficients (omitting the conurbation dummies and interactions) from the regression models. In the case of Indians and Bangladeshis, the interpretation is very simple – the proportions of adults in these categories increased by 17 and by 35 per cent respectively, across the board, with no tendency for larger or smaller increases in areas which had high densities to start with.[11] In the case of Pakistanis, Caribbeans and Africans, a slightly more complex pattern was observed – the positive coefficient on the 2001 Census term, and the negative coefficient on its square shows that the increase in each group's density was greater when the starting point was low, and less when the starting point was high.

Table 2 Regression estimates of ethnic densities by postal sector: 2007 APS outcomes predicted by 2001 Census inputs

	Indian	**Pakistani**	**Bangla-deshi**	**Caribbean**	**African**
2001 Census	**1.17**	**1.55**	**1.35**	**1.29**	**1.98**
2001 Census^2	0	**-0.71**	0.03	**-0.76**	**-4.10**
R^2	73.2%	70.0%	65.7%	62.7%	55.4%
R^2 if simple linear prediction	72.8%	69.1%	65.3%	62.0%	52.1%

Note: coefficients printed in bold type are significant at the 95% confidence level

The row of Table 2 labelled R^2 shows how much of the observed variation in 2007 outcomes is accounted for by the measure of 2001 inputs (plus the regional variables). In general area-level correlations tend to be much higher than individual-level correlations of continuous variables; on the other hand, the sampling errors in the 2007 survey data would tend to depress correlations. The accuracy of the predictions range from 73 per cent (Indians) to 55 per cent (Africans) and this gives some confidence in the predictions based on the 2001 starting point and the formulae for trends up to 2007.

Note, though, that the bottom line of Table 2 shows what the value of R^2 would have been if we had imposed a simple linear assumption on the analysis (ie omitting the term for Census 2001 squared). The prediction of 2007 densities would have been almost as accurate if no curves reflecting increasing dispersal had been allowed for.

The APS analysis of micro-data was in principle a breakthrough in the methodology of sample selections across small areas late in the 10 year Census cycle. But it turns out to demonstrate that increasing all postal sector estimates by a constant fraction within

[11] This summary is an approximation, based on assuming that changes in the composition of sectors are independent of changes in the size of sectors.

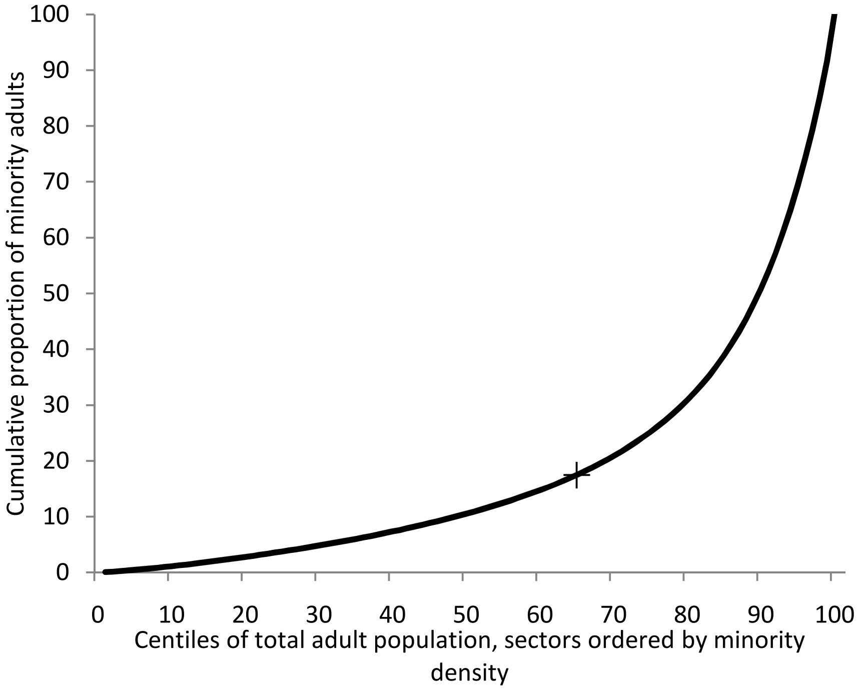
each region (derived from publicly available region-level analysis of the APS), would have been almost as accurate as fitting the slightly more sophisticated curvilinear model to small areas. This suggests that the APS sector-level analysis was not essential, though at least it confirms that applying simple growth factors was an acceptable approach.

Nevertheless, the curvilinear model (including the squared term) was slightly more accurate, and was used to establish the estimated minority ethnic densities in each postal sector.

Step 2: Exclude postal sectors with very low minority densities.

As discussed above, the distribution of minorities in Britain is a very long way from a pattern in which all minorities lived in areas where all the residents were members of minority groups. But there is still variation in densities between postal sectors, as illustrated in Figure B. Along the X axis, all the sectors in the country are ordered from the one with the lowest density to the one with the highest, and divided into 100 equal-sized groups (weighted by adult population). Up the Y axis, the proportion of ethnic minority adults is plotted cumulatively. Obviously, the very low density sectors contribute a very small proportion of the minorities, but as density increases, the curve steepens and high density sectors contribute successively a higher and higher proportion of the total, until all 100 per cent are accounted for. The plus sign (+) in the curve is plotted at the 65th percentile where still only 17 per cent of the minorities have so far been accounted for. The remaining 35 per cent of all sectors contribute the remaining 83 per cent of the target group.

Figure B Distribution of all ethnic minorities by postal sector density (2007 estimates)



It was decided to confine the ethnic minority boost sample to sectors where the estimated density exceeded 5 per cent – very close to the 65th centile used to illustrate the difference between high and low density sectors in Figure B. The average density in the selected sectors was 20 per cent. The estimated coverage for each specific group was as follows:

Indian	82%
Pakistani	93%
Bangladeshi	93%
Caribbean	86%
African	85%
Other included groups	82%

Note that the 17 per cent of members of minority groups living in the excluded low-density sectors (below 5 per cent) are still included in the survey, because of their representation in the main equal probability sample

Step 3: Estimate the number of addresses to be selected, using a sampling fraction weighted towards sectors containing the scarcest target groups.

Confining attention now to the postal sectors with estimated minority densities in excess of 5 per cent, the aim was to select proportionately large samples of addresses in areas with proportionately large numbers of members of the target minority groups.

The simple solution would have been to select a sample of addresses within each sector using a sampling fraction calculated as a function of the square root of its minority density. See Nazroo and others (2005) for a proof of this solution.

In fact the design required not a simple sample of members of all minority groups, but a structured sample in which each of the five target groups would contribute about the same number of respondents. This meant focussing the sample on areas populated by the scarcest target minority group (Bangladeshis), with relatively less emphasis on areas where the largest minority groups (Indians) lived. It is important to bear in mind that sampling fractions can vary between areas (according to the estimated composition of their populations) but cannot vary between households because at this stage we did not yet know households' ethnic composition. All the households in areas with high sampling fractions will have above-average probabilities of selection, not just members of the scarcest target groups.

The sampling fraction applied within each postal sector was as follows:

$$[\mathrm{sqrt}(0.25*\%\mathrm{Ind} + 0.50*\%\mathrm{Pak} + 10*\%\mathrm{Bang} + \%\mathrm{Carib} +\%\mathrm{Afric})]/83.95$$

where %Ind (and so on) represents the proportion of the adult population of each postal sector who were Indian (and so on). This formula is taken from Nazroo and others' (2005) outline design for the LSEM. It can be seen that the sampling fraction in an area of high Bangladeshi concentration will be 6.3 times the fraction in an area of similarly high Indian concentration (√40). The theoretical range of fractions is between about

3.8 per cent (if there was an area 100 per cent of whose residents were Bangladeshi) and 0.13 per cent (in an area 5 per cent of whose residents were Indian).[12]

The constant (83.95) was calculated to yield the correct number of issued addresses in total. This was done iteratively (at the end of the sampling procedure) with successive adjustments to the constant being made until the predicted number of achieved interviews with Bangladeshis adults reached 1,000.

At this preliminary stage, these fractions were applied to the estimated populations of each postal sector to calculate the number of addresses that would be selected. No actual selection took place.

It was also necessary to estimate the yield in terms of the number of adults in each ethnic group. These yields were based on the following assumptions:

- The number of households identified would be 90 per cent of the number of addresses issued, after allowing both for deadwood (ineligible addresses) and for multiple household addresses. This factor was based on NatCen's experience of response rates in the Family Resources Survey.
- The overall response rate (adults interviewed divided by adults in qualifying households would be 57.5 per cent for adults outside the Caribbean category, and 46 per cent for Caribbeans. (This difference was based on previous evidence that Caribbeans had below average response rates – eg Modood, Berthoud and others 1997.) These factors were designed to allow for non-response by whole households, and also for non-response by individuals within households.
- It was further assumed that response rates would vary between regions and sub-regions pro rata to response variations in the Family Resources Survey. This factor ranged between 60.0 per cent (Inner London) and 82.5 per cent (south Yorkshire).
- The calculations also required an estimate of the number of adults in each minority group as a proportion of the number of households identified as containing any of that group. These estimates, which ranged from 1.34 (Africans) to 2.23 (Pakistanis) were derived from the APS.

Note that these planning assumptions are by no means findings of the survey, They were needed simply to provide estimates of the number of addresses to be selected. To the extent that the outcomes were different from the assumptions, there will probably be a small loss of efficiency in the sample design, but no loss of representativeness in the final data.

These assumptions were used to provide a first estimate of the number of addresses that would be selected in each postal sector, of the yield within each sector in terms of interviews with households and adults in each ethnic group, and of the total yield for the sample as a whole.

[12] The range of actual sampling fractions is reported in the section on weighting, below.

Step 4: Cluster low-yield sectors

The initial first estimate of the selection procedure suggested that 44,000 addresses should be selected among the 3,145 postal sectors – an average of 14 addresses selected in each sector. After allowing both for screening and non-response, this was expected to yield interviews in 4,500 households – 1.4 households achieved per sector.

Remember that the areas included at this stage consist of all postal sectors with a minority density above 5 per cent. As explained, the sampling fraction applied within each sector was proportional to (a complex measure of) minority density – so that sectors just above the 5 per cent cut off would have very low sampling fractions. And of course sectors varied in the number of addresses available for selection. This means that some sectors were predicted to have no addresses actually selected; that many more would yield no actual interviews; and that among those yielding any interviews, there would be a wide range between sectors with one or two, and others with as many as 28. These figures are summarised in the top panel of Table 3.

Table 3 Distribution of postal sector sample sizes, as estimated before and after clustering.

		Mean per sector	Number of zeroes	Maximum
Initially	**3145 sectors**			
	Addresses issued	14.1	99	115
	Household interviews	1.4	1645	28
After clustering	**771 sectors**			
	Addresses issued	56.8	None	309
	Household interviews	6.0	10	28

This initially-calculated plan was inefficient. From the point of view of organising the screening operation, a very large number of assignments would be issued with the expectation of a zero outcome. From the point of view of organising the interviewing in the second and subsequent waves, a large number of assignments would contain only one or two participating households.

So it was decided to sub-sample sectors with very small expected assignments – reducing the number of such sectors, and conversely increasing the size of the assignments in each remaining small sector. This procedure can be thought of as equivalent to the technique of 'clustering' in two-stage random samples.

- The sectors were divided into four groups: those with predicted yields of 0, 1, 2 or 3-plus households.
- Within each group, they were ordered geographically by region, postal area and postal district.
- Within each group, 1 in 16 were retained from among sectors with a predicted yield of 0 households, 1 in 8 from among those with a predicted yield of 1; 1 in 4 from among those with a predicted yield of 2.
- The sampling fraction for selecting addresses within the retained low-yield sectors was multiplied by 16, 8 or 4, as appropriate, so the overall distribution of selected addresses across small and large sectors was self-weighting. That

is, if (say) 6 addresses had originally been expected to be selected in a retained sector with a predicted yield of one household, 8*6=48 addresses would now be selected, and the predicted yield rises to 8.

- All of the sectors with originally predicted yields of three or more households were retained, with no adjustment to their within-sector sampling fractions.

This procedure for reducing the number of small assignments and increasing their size produced a revised profile of sectors as described in the lower panel of Table 3. The total number of selected sectors fell from 3145 to 771. Only ten of these would now be predicted to yield no household interviews, and the average number of interviews per sector was six. Most (78 per cent) of the total number of expected interviews would still take place in areas with an original expectation of three or more responding households, and so were not affected by the clustering procedure.

In summary, the boost sample can be thought of as consisting of five primary strata:

- The two-thirds of all postal sectors with minority densities below 5 per cent not included in the boost sample at all (though they are covered by the main sample).
- Postal sectors with initially-predicted yields of zero households, which were sub-sampled at the rate of 1 in 16.
- Sectors with initially-predicted yields of one household, which were sub-sampled at the rate of 1 in 8.
- Sectors with initially-predicted yields of two households, which were sub-sampled at the rate of 1 in 4.
- Sectors with initially predicted yields of three or more households, all of which were included in the boost sample.

Step 5: Select addresses

The sample at this stage consists of a list of 771 postal sectors, each with a sampling fraction.

Where more than 100 addresses would be selected for screening in any single sector, the address list was split in half to make two assignments each of less than 100. Where the initial selection would be more than 200, it was split in three. Sectors with very small expected selections were grouped with near neighbours to impose a minimum assignment size.[13] This splitting and combining of assignments was designed to improve the efficiency of fieldwork allocations - it has no effect on the theory or structure of the sample as a whole.

[13] A PSU (as originally selected) was defined as 'small' if either:
(a) the expected number of issued addresses was less than 15; OR
(b) the expected number of achieved interviews was less than 2 and the expected number of issued addresses was less than 50.

A 'small' PSU could be merged with a neighbouring PSU if their centres (defined by O/S reference) were within 15km of each other.

Assignments (as defined in the previous paragraph) were then allocated at random into 24 equal groups, to be covered in each of the 24 months of the wave 1 fieldwork period (January 2009-December 2010)..

Addresses were selected from each sector's PAF address lists in the normal way, applying the fractions laid down by the previous steps to the list of addresses stratified in postcode order, from a random start number.

A review of the outcome of the first five months of fieldwork suggested that the number of interviews with Bangladeshis was lower than expected, and that it would fall short of the 1,000 target over the two year period. It was therefore decided to issue more addresses in areas of very high Bangladeshi concentration. In each of the 18 sectors with the highest concentration, additional addresses were selected, 1½ times as many as had been selected in the first draw (so that the total issued addresses in those sectors was 2½ times the original figure). These extra assignments were allocated at random across the remaining 12 months of fieldwork, January-December 2010.

Step 6: Screening

For each assignment (as defined interviewers would visit every address. The task was to ask questions of one adult member of the household about the ethnic background of all members of the household. The screening question was listed on page 6, and full copy of the screening question sequence is in Appendix 1. If any household member was reported to be in any of the minority groups listed, the household was retained to the next step; otherwise (for example if all residents were white), the household was rejected from the sample. This is known as the primary screening.

The selection of addresses in postal sectors was designed to achieve 1,000 Bangladeshi adults, and *at least* 1,100 or 1,125 adults in the other four included minority groups. This unavoidably yielded more than these minima in some groups. The objective of selecting a series of similar boost samples of specific minority groups within the constraints of the boost budget, led to the decision to limit the size of each of these group-samples to 1,100 or 1,125. If more than that number of adults were (expected to be) identified, then a secondary screening procedure was required to achieve the target number of interviews. The secondary screening was designed to:

- Achieve 1,000 interviews with Bangladeshi adults
- Interview as many 'mixed-Indians' and 'mixed-Caribbeans' as could be identified.
- Achieve 1,125 interviews with Indians and with Caribbeans, including the 'mixed' groups in the target total.
- Achieve 1,100 interviews with Pakistanis, and with Africans
- Interview as many Sri Lankans, Chinese and Turks as could be identified. These were the three non-target minority groups where the roughly estimated numbers were potentially large enough for country-of-origin specific analysis to be possible.
- Achieve 1,100 interviews with 'other included groups', including Sri Lankans, Chinese and Turks in the total.

These considerations led to the retention fractions reported in Table 4.

Although these secondary sampling fractions were designed to achieve target numbers of individual adults, they were applied to whole households. The field procedure applied the fractions hierarchically: if any member of a household was in an ethnic group requiring 100% selection, then that household, and all its members, were recruited to the survey. If a household failed that test, but any member was in an ethnic group requiring a 91 per cent fraction, then that probability was applied to the household and all its members. And so on, down from the higher to the lower fractions in sequence.

Table 4 Secondary screening retention fractions

Main category	Sub-group	Fraction for assignments issued from January to December 2009	Fractions for assignments issued from January 2010
Indians	Indians (non-mixed)	50%	65%
	mixed Indians	100%	100%
Pakistanis	Pakistanis	54%	100%
Bangladeshis	Bangladeshis	100%	100%
Caribbeans	Caribbeans (non-mixed)	91%	100%
	mixed Caribbeans	100%	100%
Africans	Africans	82%	100%
Other included groups	Sri Lankans	100%	100%
	Chinese	100%	100%
	Turks	100%	100%
	far eastern	30%	100%
	middle eastern	30%	100%

The first numerical column of Table 4 shows the secondary selection fractions applied in the first year of the survey (January-December 2009), derived directly from the calculations (based on the Census and the APS) that had been used to design all other elements of the sample. Analysis of the first five months of fieldwork returns suggested that the number of adults being interviewed in several minority groups was likely to fall below target if these secondary selections continued unaltered for the full 24 month period. For the 12 fieldwork periods January-December 2010, the secondary selection fraction for Indians was raised to 65 per cent, and all other groups were included in the survey without further screening. These revised fractions are recorded n the second numerical column of Table 4.

Step 7: Following rules

The ethnic minority panel consists of all individual members of the target and included minority groups identified at the screening stage, as described above. These are referred to as original sample members (OSMs). In households where members of minorities live with white people (or other excluded groups), the members of included minorities are defined as OSMs while the white (excluded) members of these mixed households are defined as temporary sample members (TSMs).

The following rule is that in the second and subsequent waves, all OSMs are interviewed, even if they have split up and live separately from each other, and/or with others. OSMs who are children are also followed and are administered the youth questionnaire once they turn 10 and the full questionnaire once they turn 16. Children born to OSM mothers subsequent to wave 1 also become OSMs themselves.

When original sample members (OSMs) start to live with someone who is not an OSM, all these co-residents are also interviewed, to provide information about the household in which the OSM lives. But these are temporary sample members (TSMs), included in the survey only as long as they live with an OSM. TSMs are not followed if they cease to be co-resident with any OSMs.

One apparently anomalous case is worth noting – where a household contains a child who is a member of an included minority ethnic group, but all the adults are white (or from excluded groups). (This could occur, for example, if an adopted child is attributed to the ethnic group of his/her natural parents, or if a white lone mother lives with a child whose father was from a minority group.) In this case the minority child is an OSM and should be followed (even though s/he may be too young to be interviewed), and the white adult(s) are TSMs and should be interviewed only as long as they live with the child. So there will be a few households which contribute to the minority sample even though all the people directly contributing data are white.

Step 8: Weighting

The sampling procedure described here resulted in selection probabilities of individuals that varied for three reasons:

- The fraction of addresses selected for screening in each postal sector is a complex function of the ethnic profile of the people living there (as calculated at Step 3 and applied at Step 5).
- The proportion of postal sectors sampled varied between the four strata (step 4); But note that the fraction of addresses to select in each postal sector was adjusted to reflect the sub-sampling of sectors at step 4, so differences between the four strata will cancel out.
- The fraction of households selected to take part in the full survey after initial screening varied according to the ethnic groups of the household members (Step 6)

When the boost sample is analysed, it will be necessary to apply weights to compensate for the fact that some types of people had a much higher probability of being drawn into the sample than others. The boost sample design weights (ie multiplying by the reciprocal of the calculated sampling fractions), are an intrinsic component of the design.

A more comprehensive set of weights will eventually be calculated, to take account of:

- The actual (rather than the predicted) yield of interviews in each postal sector.[14]
- The actual (rather than the predicted) number of households and of adults screened in each group.[12]
- Possible response bias (as evidenced for example by differences between the composition of the sample and the profile of the population it aims to represent, or by non-response by individuals within households).
- Weights needed to combine the boost sample with the main sample for analysis of all ethnic minorities, and indeed for analysis of the whole population.

These final weights will be calculated after the first wave of interviewing has been completed, and a separate working paper will describe and assess them in detail.

The boost sample design weights are discussed here because the variable fraction design that has been described in this paper affects the efficiency of the sample. All the estimates here are based on the predicted outcome of the survey, not the actual outcome.

The initial focus is on the sampling fractions specified for selecting addresses within postal sectors, ignoring (for the moment) the secondary selection fractions imposed on households at the screening stage. The total predicted boost sample size was 6,528 minority adults. Weights (calculated as the reciprocal of the within-sector sampling fraction) averaged 151.[15] They ranged from 858 to 34; expressed as a ratio to the average, the range was from 5.68 to 0.23. The coefficient of variation was 59 per cent. (In general, a high coefficient of variation, indicating a wide range between weights, would be expected to reduce statistical efficiency.)

Table 5 summarises the impact of these weights on the overall sample, and on the main sub-groups of ethnic minorities. Although weighting is required to counteract known variations in selection probabilities, it is not a costless exercise. Sampling errors for a weighted sample are larger than those for an equivalent unweighted sample – the wider the range of weights, the greater the increase in sampling errors Kish 1992). This effect can be presented in terms of the size of a hypothetical unweighted sample that would be as accurate as the actual weighted sample – recorded as the 'effective sample size' in Table 4. This is calculated as:

$$(\sum W)^2/\sum(W^2)$$

where W is the weight assigned to each case.

[14] The actual yields will vary from the predicted yields as a result of sampling error, variations between predicted and actual densities in the sample sectors; variations in identification rates at the screening stage, and variations in response rates.

[15] The weights reported here are based on the initial sampling fractions, and take no account of the additional addresses selected in high-density Bangladeshi areas in the second year of fieldwork (se page x)

Table 5 Predicted impact of variation in selection probabilities on precision

	Predicted number of adults interviewed	Coefficient of variation of postal sector weights	Effective sample size	Efficiency
Overall	6528	59%	4843	74%
Indian	1125	51%	893	79%
Pakistani	1100	58%	827	75%
Bangladeshi	978	68%	669	68%
Caribbean	1125	50%	901	80%
African	1100	51%	875	80%
Other included groups	1100	57%	832	76%
Target groups	5428	59%	4041	74%

For the boost sample as a whole, the 6,528 predicted interviews with minority adults are estimated to be as accurate as an unweighted sample of 4,843. The final column of Table 5 interprets this as an efficiency rate of 74 per cent (4843/6528).

The strategy behind the sample design was that analysis of all ethnic minorities combined was not especially important, and in any case would be based on a very large sample size. The key objective was to enable analysts to look separately at each of the target groups. Table 5 shows that the predicted samples of 1,000 Bangladeshis and 1,100 or 1,125 in each of the other groups are expected to produce effective samples ranging from 669 to 901, at an efficiency ranging from 68 per cent to 80 per cent. The lowest efficiency rating is for Bangladeshis – this is an inevitable outcome of the fact that the sample design had to be twisted hardest to produce an adequate sample of this group, because it is the smallest of the target communities.

The calculations in Table 5 are based entirely on the weights required to counteract variations in the within-sector sampling fractions. A second weight, applied together with the first, is required to counteract variations in the secondary selection of households at the screening stage. This second weight does not have much bearing on the effective sample size of each target minority group (because all or most of the members of each group have the same selection fraction), but will further reduce the accuracy of estimates based on all minorities together. Focussing on the combination of the five target groups,[16] Table 5 shows that 5,428 actual interviews are equivalent to 4,041 after the sector weights are applied. Further calculations (not shown in the table) suggest that the effective sample size reduces to 3,578 when the year 1 screening-selection weights are applied as well – an overall efficiency rate reduced to 66 per cent.[17]

[16] The 'other included group' are left out of the calculations here, because we have no firm basis for estimating the numbers of Sri Lankans, Chinese and Turks assigned a 100 per cent selection, and the far eastern and middle eastern people assigned a 30 per cent rate.

[17] The figures in this paragraph are based on the secondary selection fractions set at the start of the survey, and applied between January and December 2009 – see the first column of Table 3.

Discussion

This project can be thought of as a double extension to the social research evidence base.

- In the context of longitudinal studies, it provides the largest ever panel sample of members of minority ethnic groups, enabling much more detailed analysis than was possible with the BHPS.
- In the context of ethnicity research, it provides significant panel data for the first time, enabling much clearer interpretation of family dynamics and causal inferences than has been possible with previous cross-sectional surveys..

The objective was to combine the minority respondents identified in the boost with those interviewed as part of the main equal probability sample, to provide evidence about variations between minority ethnic groups, and between them and the white majority. On the whole, the questions asked of both samples were the same, although there was a small section of questions reserved for the boost.

The techniques adopted to select the sample have much in common with previous national samples of minority ethnic groups. But five points are worth highlighting.

- *Use of the APS to update census estimates*: Although the Census provides detailed and accurate data about the ethnic composition of small areas, the information becomes out of date as the ten year period between Censuses elapses. We used micro-data from the 2007 Annual Population Survey to derive estimates of the rate of change in ethnic composition at different levels of (2001) density - separately for each of our target groups. This allowed us to estimate the current composition of each postal sector. Although it turned out that these estimates were not markedly different from what would have been expected if average changes at regional level had been applied at sector level, the method at least confirmed that there had been no major change in residential patterns. The method could easily be replicated for other samples of ethnic minorities, and indeed for other types of sub-group.
- *Separate targets for five specific minority groups*: Previous surveys have almost always sought samples of ethnic minorities, thought of as a single group. But theoreticians and analysts have been emphasising ethnic diversity - the importance of variations between minority groups. For the first time, the boost sample for *Understanding Society* targeted five specific minority groups, seeking approximately equally sample sizes for each one.
- *Formula to derive within-sector sampling fractions*: The standard approach is to identify a stratum of areas with high minority densities, and apply a higher (or additional) sampling fraction to all areas within that stratum. With five distinct targets, a more complex method was needed, which hyper-sampled areas where the scarcest minorities lived, and only slightly over-sampled areas where the most common minorities lived. A formula was applied (page 14), which expressed the composition of each postal sector as a function of the five group densities. Sampling fractions varied continuously across all sectors – in effect, each sector was treated as a separate stratum.
- *Clustering*: Standard sampling procedures often involve a two stage design, in which a sample of small areas is selected, and then samples of addresses within those sectors. The technique adopted on this occasion was to estimate the expected

yield of household interviews in every postal sector (whose overall minority density was at least 5 per cent). All sectors where at least three household interviews were predicted were retained in the sample (and these represent the majority of households in the boost). Where 0, 1 or 2 household interviews were predicted, a sub-sample of sectors was selected, but the sampling fractions within the retained sectors were increased to compensate.

- *Variable screening procedure*: An efficient sample of minority groups requires that households which do not contain members of the target group are screened out at the fieldwork stage. This procedure was incorporated in the current design, using a specially drafted question sequence. But because the aim was to achieve approximately equal numbers of interviews in each of five target groups, it was appropriate to introduce secondary screening, so that some households containing the most common minority groups (especially Indians) were deselected at random to limit the size of that sub-sample.

Although data on the outturn of the early months of fieldwork has been used to make some fine adjustments to the sampling procedures (see page 17 and page 19), this working paper has mainly described the design of the boost sample, effectively from the point of view available prior to the launch of the survey. Another working paper will report on and asses the outcome of the sample, and draw further lessons for future studies.

References

Brown, C. (1984), *Black and White Britain*, Heinemann

Brown, C. and Ritchie, J. (1982) ’Focussed enumeration: the development of a method for sampling ethnic minority groups’, Policy Studies Institute/Social and Community Planning Research

Daniel, W.W. (1967) *Racial Discrimination in England*, Penguin

Duncan, O. and Duncan, B. (1950) *The Negro Population of Chicago*, University of Chicago Press

Kish L (1992) ‘Weighting for unequal P’, *Journal of Official Statistics* 8, 183-200

Lynn, P (2009) ‘Sample design for *Understanding Society*’, *Understanding Society* Working Paper 2009-01, University of Essex

Modood, T., Berthoud, R., Lakey, J., Nazroo, J., Smith, P., Virdee, S. and Beishon, S. (1997) *Ethnic Minorities in Britain: diversity and disadvantage*, Policy Studies Institute

Nandi, A., Platt, L. and Burton, J. (2008) ‘Who are the UK’s minority ethnic groups? Issues of identification and measurement in a longitudinal survey’. ISER Working Paper 2008-26, University of Essex

Nazroo, J., Berthoud, R., Erens, B. Karlsen, S. and Purdon, S (2005) ‘A Longitudinal Survey of Ethnic Minority People: focus and design’ University College London

Peach, C. (1996) ‘Does Britain have ghettos?’, *Transactions of the Institute of British Geographers*, vol 21, no 1

Owen, D. and Green, A. (2003) ‘A Scoping Study for a Longitudinal Survey of Ethnic Minorities for the UK’, University of Warwick

Smith, D.J. (1976) *Racial Disadvantage in Britain*, Penguin

Smith, P, Pickering, K., Williams J. and Hay, R. (2010) 'The efficacy of focused enumeration' paper to Royal Statistical Society, May 2010

Appendix: the screening question sequence

(see next page)

D.7 SHOW SCREENING CARD

Does anyone living at this address come from, or have parents or grandparents from **any** of the following ethnic groups?

CODE ALL THAT APPLY

Indian	01	**Go to D.8**
Mixed Indian – (parents or grandparents from Indian ethnic group **AND** parents or grandparents from a non-Indian ethnic group)	02	
Pakistani	03	
Bangladeshi	04	
Sri Lankan	05	
Caribbean / West Indian	06	
Mixed Caribbean/West Indian (parents or grandparents from Caribbean/West Indian ethnic group **AND** parents or grandparents from a non-Caribbean/West Indian ethnic group)	07	
North African	08	
Black African	09	
African Asian	10	
Chinese	11	
Far Eastern (includes Filipino, Thai, Malaysian, Japanese, Vietnamese, Singaporean, Indonesian, Korean, Burmese)	12	
Turkish	13	
Middle Eastern / Iranian (includes Israeli, Palestinian, Lebanese, Syrian, Jordanian, Yemeni, Saudi, Iraqi, Afghani, other Gulf states)	14	
None of these	96	**Go to F.6 (code 770)**
Unable to complete screening questions	95	**Go to F.7**

D.8 INTERVIEWER: IF CODE 1 OR CODE 6 AT D.7, CHECK THAT ALL PARENTS AND GRANDPARENTS ARE FROM INDIAN (CODE 1) OR CARIBBEAN/WEST INDIAN (CODE 6) GROUPS. IF NOT USE CODE 2 FOR MIXED INDIAN OR CODE 7 FOR MIXED CARIBBEAN/WEST INDIAN AS APPROPRIATE.

D.9 Does anyone living at this address come from, or have parents or grandparents from **any** of the following ethnic groups?

CODE FROM D.7

Mixed Indian – (parents or grandparents from Indian ethnic group **AND** parents or grandparents from a non-Indian ethnic group)	02	**Go to E.1**
Bangladeshi	04	
Sri Lankan	05	
Mixed Caribbean/West Indian (parents or grandparents from Caribbean/West Indian ethnic group **AND** parents or grandparents from a non-Caribbean/West Indian ethnic group)	07	
Chinese	11	
Turkish	13	
None of these	96	**Go to D.10**

D.10 Transfer eligibility number from front of ARF □□ and then code

Eligibility number <=30	1	**Go to D.11**
Eligibility number in range 31-50	2	**Go to D.12**
Eligibility number in range 51-54	3	**Go to D.13**
Eligibility number in range 55-82	4	**Go to D.14**
Eligibility number in range 83-91	5	**Go to D.15**
Eligibility number >=92	6	**Go to F.6 (code 770)**

D.11 Does **anyone** living at this address come from, or have parents or grandparents from any of the following ethnic groups?

CODE FROM D.7

Indian	01	**Go to E.1**
Pakistani	03	
Caribbean / West Indian	06	
North African	08	
Black African	09	
African Asian	10	
Far Eastern (includes Filipino, Thai, Malaysian, Japanese, Vietnamese, Singaporean, Indonesian, Korean, Burmese)	12	
Middle Eastern / Iranian (includes Israeli, Palestinian, Lebanese, Syrian, Jordanian, Yemeni, Saudi, Iraqi, Afghani, other Gulf states)	14	
None of these	96	**Go to F.6 (code 770)**

D.12 Does **anyone** living at this address come from, or have parents or grandparents from any of the following ethnic groups?

CODE FROM D.7

Indian	01	**Go to E.1**
Pakistani	03	
Caribbean / West Indian	06	
North African	08	
Black African	09	
African Asian	10	
None of these	96	**Go to F.6 (code 770)**

D.13 Does **anyone** living at this address come from, or have parents or grandparents from any of the following ethnic groups?

CODE FROM D.7

Pakistani	03	**Go to E.1**
Caribbean / West Indian	06	
North African	08	
Black African	09	
African Asian	10	
None of these	96	**Go to F.6 (code 770)**

D.14 Does **anyone** living at this address come from, or have parents or grandparents from any of the following ethnic groups?

CODE FROM D.7

Caribbean / West Indian	06	**Go to E.1**
North African	08	
Black African	09	
African Asian	10	
None of these	96	**Go to F.6 (code 770)**

D.15 Does **anyone** living at this address come from, or have parents or grandparents from **Caribbean/West Indian** ethnic group?

CODE FROM D.7

Yes	1	**Go to E.1**
No	2	**Go to F.6 (code 770)**