



Home Office

A feasibility study for a survey of migrants

**Dr Patten Smith, Andrew Cleary, Meghann Jones, Sheonaidh Johnston
and Pamela Bremner (Ipsos-Mori)**

Dr James Brown and Professor Richard Wiggins (Institute of Education)

**Report prepared by Ipsos MORI and the Institute of Education
for the UK Border Agency**

January 2011

Occasional Paper 92

This report was commissioned in January 2009

The views expressed in this report are those of the authors, not necessarily those of the Home Office (nor do they reflect Government policy).

You may re-use this information (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence.

To view this licence, visit <http://www.nationalarchives.gov.uk/doc/open-government-licence/> or write to the Information Policy Team, The National Archives, Kew, London TW9 4DU, or e-mail: psi@nationalarchives.gsi.gov.uk

© Crown copyright 2011

ISBN 978 1 84987 398 7

Overview

The Analysis, Research and Knowledge Management Directorate (ARK) in the UK Border Agency commissioned Ipsos MORI and the Institute of Education to undertake a feasibility study for a large-scale face-to-face survey of migrants in the UK. The purpose of the feasibility study was to inform the design of a survey of migrants, looking at how to design and interview a statistically robust sample and the appropriate questions to adopt in such a survey.

The feasibility study involved:

1. workshops with the UK Border Agency and other stakeholders to identify survey requirements;
2. the development of a definition of 'migrant' for use in the survey;
3. the development of a sample design;
4. testing fieldwork recruitment and data collection methods to be used; and
5. the development of a questionnaire, including cognitive testing in several languages.

This report focuses on the sampling and field methods aspects of the feasibility study. A separate technical report provides details on other issues, including question development.

To inform the development of the survey methodology and the survey questionnaire discussion groups and depth interviews were undertaken with representatives of economic migrant communities and individuals working with refugees and asylum seekers in England and Scotland. Key issues surrounding survey objectives, terminology, and field methods were identified.

A useable survey population definition was developed after stakeholder consultation such that, for the purpose of this study:

'A migrant is defined as someone who arrived in the UK in 1990 or later, who was a non-UK national on entry, whose usual place of residence prior to entry was not in the UK, and who has lived in the country for at least three months (one month if an asylum seeker or refugee).'

It was concluded that different types of migrants will require different sampling and screening methods. Asylum seekers could be sampled directly from administrative records but other migrant types would have to be screened in the field. Small area estimation methods were successfully used to identify areas in which migrants are more heavily concentrated in order to inform the sample design.

A number of field screening methods were investigated and field tested. Despite testing some innovative methods, traditional screening methods were found to still have an important role to play. It was concluded that non-asylum-seeker migrants should be screened through a combination of traditional screening methods for areas of higher migrant concentration

and an adapted form of the Waksberg method (a two-stage method originally developed to minimise the number of calls made to ineligible numbers in telephone surveys) for areas of lower migrant concentration.

An overall approach to a sample design was outlined and illustrative designs were presented applicable potentially to 6,000 interviews in England and 3,000 in Scotland. It was concluded that the survey should aim for 80 per cent coverage of the migrant population and that fieldwork effort should be disproportionately concentrated in areas of higher migrant concentration.

The overall conclusion from the work was that a high quality survey of migrants would be feasible although a detailed design would need to take into consideration the level of funding available and how competing design requirements should be prioritised.

In carrying out any such survey a large-scale dress rehearsal survey should be conducted before resources are fully committed in order to confirm the assumptions underlying the sample design and to field test the questionnaire and field translation/interpreting procedures.

Contents

Executive summary	v
1. Introduction	1
2. Survey delivery: key outcomes from the migrant workshops	3
3. Defining the survey population of migrants	6
4. Sampling and field methods	8
4.1 Investigation of potential sampling frames	8
4.2 Possible sampling methods for a survey of migrants	16
4.3 Towards a sample design	24
Appendix 1 Sampling asylum seekers	32
Appendix 2 Approaches to sampling that rely on social networks	36
Appendix 3 Approaches to sampling that rely on localised clustering	40
Appendix 4 Glossary	44
References	45

List of figures and tables

Table 4.1	Modelling the relationship between administrative sources and the APS at Local Authority level	14
Table 4.2	Adaptive cluster sampling field test outcomes	20
Table 4.3	Concentration and coverage of migrant households (England)	26
Table 4.4	Concentration and coverage of migrant households (Scotland)	27
Table 4.5	Sample sizes and cost	29
Table A1.1	Results from asylum field tests	34
Table A2.1	Multiplicity sampling outcomes	37
Table A2.2	Outcomes for new sample members from multiplicity sampling	39
Table A3.1	Outcome from Adaptive Cluster Sampling	42
Table A3.2	Adaptive cluster sampling screening rates in original addresses and newly generated addresses	43
Figure 4.1	Sample structure for England	31
Figure A3.1	Example of adaptive cluster sampling	41

Executive summary

Introduction

The Analysis, Research and Knowledge Management Directorate in the UK Border Agency commissioned Ipsos MORI and the Institute of Education, via competitive tender, to undertake a feasibility study for a large-scale face-to-face survey of migrants in the UK. A survey of migrants may help fill in a number of evidence gaps important for policy development for the UK Border Agency and other government departments which might only be filled by surveying migrants themselves:

- why migrants choose to come to the UK and why they might leave or settle permanently;
- what migrants and their dependants are doing in the UK (e.g. work, education);
- social impacts of migration, including the use of public services and integration;
- how migrants are contributing to the UK;
- experiences of the immigration system; and
- perceptions of the UK and local areas.

Current data sources on migrants are limited in both the type of information held and on whom it is held. Current surveys do not collect all of this information from the current stock population of migrants and administrative data systems do not record information that is sufficiently rich to inform many policy areas.

The purpose of the feasibility study is to inform the design of a survey of migrants. It comprises five interdependent modules:

1. workshops with the UK Border Agency and other stakeholders to identify survey requirements;
2. the development of a definition of 'migrant' for use in the survey;
3. the development of a sample design;
4. testing fieldwork recruitment and data collection methods to be used; and
5. the development of a questionnaire, including cognitive testing in several languages.

The work was undertaken between January 2009 and April 2010. This report, which focuses on the sampling and field methods aspects of the feasibility study, details:

- survey delivery, key outcomes from migrant workshops;
- defining the survey population of migrants; and
- sampling and field methods.

A separate technical report provides details on question development and additional interim reports. The questionnaire developed as part of the feasibility is also published separately.

Survey delivery: Key outcomes from the migrant workshops

To inform the development of the survey methodology and a possible survey questionnaire, one discussion group and seventeen depth interviews were undertaken with representatives of economic migrant communities from a range of countries within and outside the European Union (EU), and with individuals working with refugees and asylum seekers in England and Scotland.

The discussions raised a number of important issues in relation to methodology and questionnaire design. Key issues identified were as follows.

- That official definitions of a migrant and concepts of immigration status may not match migrants' own perceptions of their status and indicates a need to capture both actual status (for example, nationality listed on passport or legal visa category) and perceived status.
- The terms 'visitors' or 'foreign workers' were preferred to the term 'migrants'.
- Terminology surrounding migration was seen to be contentious and will require sensitive handling in any survey. The term 'immigration status' should probably be avoided and the term 'refugee' was preferred to 'asylum seeker' (although the differences between the two terms were recognised).
- There was considerable support for a survey of migrants, but presentation of the survey was a concern, particularly if presented as a UK Border Agency survey. Engagement with community organisations/charities would be key to any survey's success in order to present the survey as bona fide, raise awareness and encourage participation. A cold-calling approach for asylum seekers may have adverse consequences for response rates in this group.

Defining the survey population

The population definition for the survey was primarily guided by the expressed policy interests of stakeholders but also gave consideration to definitions already in use.

For the purposes of this study a survey population definition was agreed as:

'A migrant is defined as someone who arrived in the UK in 1990 or later, who was a non-UK national on entry, whose usual place of residence prior to entry was not in the UK, and who has lived in the country for at least three months (one month if an asylum seeker or refugee).'

Although this includes some groups of lesser interest, it was felt that this was preferable to a more complex definition excluding them, and the definition was felt to have considerable practical advantages in developing an appropriate methodology over other definitions considered.

Investigation of potential sampling frames

Two uses of a sample frame were identified:

1. as a direct source of sample members;
2. as a means to identifying areas in which migrants live and in which fieldwork could be concentrated.

Investigation of potential frames for a survey of migrants showed that frame information was only sufficiently up to date to allow for direct sampling for asylum seekers. A small-scale field test concluded that with appropriate fieldwork procedures, direct sampling of this group would prove feasible.

However, a number of other frames were considered appropriate for identifying small areas (e.g. wards) where migrants were present. This information would be used to inform the sample design.

Small area population estimation techniques were used to produce updated 'size' indicators for wards based on the current Annual Population Survey (APS) data at the Local Authority (LA) level and other indicators at the ward level. A Local Authority level regression model was set up which defined parameters linking the patterns of LA level administrative data to patterns seen in LA APS estimates. A synthetic assumption was then made that the fitted LA level patterns could be applied at the ward level to get 'predictions' of migrants at that level. In this way, ward level predictions depended upon wards' specific profiles with respect to the administrative data. The following variables were used in the final model.

- Annual Population Survey (APS) estimates of those 'not born in the UK' restricted to those individuals that arrived since 1990.
- The 2001 Census count of those 'not born in the UK'.
- National Insurance Number (NINo) data split into whether the individual was from an EU15 country or a new EU (A8/A2) country.
- Data from the UK Border Agency on the number of individuals granted 'leave to remain' (LTR) within a year.
- Data from the UK Border Agency on the number of individuals granted 'indefinite leave to remain' (ILR) and those granted Citizenship within a year.
- Interactions to allow for a differing relationship within Scotland.

The most important predictor variables came from the Census count, UK Border Agency data (grants of leave and indefinite leave to remain and citizenship), and NINo EU15 data. As part of a further project (outlined in the separate technical report) local experts commented on the local patterns produced from the modelling. Broadly speaking they were supportive of the patterns found by the model.

The authors recommend that the same approach should be taken in any survey of migrants using updated administrative data.

Possible sampling methods for a survey of migrants

The authors investigated a number of possible methods for sampling and identifying migrants in the field for use when direct sample frames are unavailable. They summarise the conclusions of this work below.

1. The authors rejected quota sample approaches because these would not allow robust estimation of key survey parameters.
2. If it proved possible to follow up migrants identified by the Labour Force Survey (LFS), this would prove to be a cost-effective method for identifying a high quality sample of migrants. It would, however, extend the fieldwork period and might limit coverage of very short-term migrants. There may also be issues surrounding data access and permissibility of recontacting LFS respondents.
3. A conventional screening approach involving in-field eligibility checks at large samples of addresses (focused in high concentration areas) would almost certainly be a major component of a migrant survey. These methods would, however, prove especially costly in areas of lower migrant concentration.
4. Field tests showed that methods for focusing screening based upon social networks such as multiplicity sampling would be infeasible in practice.
5. Methods based upon local clustering are more promising. Field tests of adaptive cluster sampling (ACS) indicated that this could be made to work in the field, but that it would prove hard to control sample size and structure.
6. An adapted form of the Waksberg sampling method (a two-stage method originally developed to minimise the number of calls made to ineligible numbers in telephone surveys) was considered the most promising method for use in areas of lower migrant concentration.
7. Field tests indicated that it should be feasible to sample asylum seekers direct from UK Border Agency administrative records.

Towards a sample design

A general approach to the sample design was proposed. This would involve restricting coverage of the migrant population to the 80 per cent living in more concentrated wards (identified through small area estimation methods) and therefore excluding areas with very low numbers of migrants. A two-stage sample of addresses would be selected in these areas using wards as primary sampling units.

Within the higher migrant concentration wards (eg where 10% or more of the population were eligible migrants) conventional face-to-face screening would be carried out. An adapted Waksberg method would be used in the remaining, lower concentration, wards.

Illustrative sample designs delivering 6,000 interviews in England and 3,000 in Scotland were outlined. On this basis, and assuming that an equal probability address sampled was drawn, it was concluded that around 82,500 addresses would need to be screened in England and 125,000 in Scotland.

The number of addresses required for screening (and hence costs) could be reduced considerably by sampling addresses at higher rates in more concentrated areas than in less concentrated areas. Although this approach would decrease the precision of survey estimates, it was concluded that it would be more cost-effective than an equal probability approach and is therefore the authors' recommended approach.

Next steps

The main conclusion to be drawn from this work is that a high quality survey of migrants would be feasible, although it would be necessary to test a detailed design, taking into consideration available levels of funding, and to consider how competing design requirements should be prioritised. The authors strongly recommend that a large-scale dress rehearsal survey be conducted before resources are fully committed to a main survey in order to establish how realistic the authors' assumptions about local area concentrations of migrants are; to establish likely levels of response rates; to test the efficacy of the proposed adapted Waksberg method; to field test the questionnaire on a reasonably large random probability sample of migrants; and to test the efficacy of field translation/interpreting procedures.

1. Introduction

The Analysis, Research and Knowledge Management Directorate in the UK Border Agency commissioned Ipsos MORI and the Institute of Education, via competitive tender, to undertake a feasibility study for a large-scale face-to-face survey of migrants in the UK. A number of significant methodological challenges had to be addressed by the feasibility study with a view to ensuring that any survey of migrants would be implemented successfully. Indeed, a migrant survey may require the use of methodologies that have not hitherto been used in large-scale social surveys in the UK.

A survey of migrants might fill in a number of evidence gaps which can only be filled by speaking to migrants directly. These are:

- why migrants choose to come to the UK and why they might leave or settle permanently;
- what migrants and their dependants are doing in the UK (e.g. work, education);
- social impacts of migration, including the use of public services and integration;
- how migrants are contributing to the UK;
- experiences of the immigration system; and
- perceptions of the UK and local areas.

Current data sources on migrants are limited in both the type of information held and on whom it is held. For example, although the Labour Force Survey has good coverage of the private residential household population and hence of migrants who are part of that population, it collects little direct information on migrants and focuses almost exclusively on collecting data relating to the labour market. The International Passenger Survey (IPS) collects information at the time of arrival of new immigrants, and, as such, collects nothing on their later status and activities in the UK. Administrative records do not collect information that is sufficiently rich to inform many policy areas even though it is likely that it has good coverage of some migrant groups.

Evidence from a survey of migrants might be relevant to the UK Border Agency's policy development, particularly in relation to the Points-Based System for non-EU migrants, but also to understand drivers of migration for EU migrants (including those from new Accession countries) who are free to come and work in the UK. Evidence from a survey of migrants might also inform a debate regarding both the social and economic impacts of migration.

This approach to a survey of migrants aims to encourage joined-up approaches across-Government to policy making on migration based on a consistent evidence base. The feasibility study was supported by a number of other government departments as well as the UK Border Agency and co-funded by Communities and Local Government (CLG), Department for Work and Pensions (DWP), the Scottish Government, the Home Office and the Migration Advisory Committee. A number of government departments and devolved administrations participated in a workshop in January 2009, at the start of the feasibility study, to establish the key requirements across government for such a survey.

The purpose of the feasibility study was to inform the design of a large-scale survey of migrants. Key issues addressed by the feasibility study included how a 'migrant' is to be defined and how migrants are to be sampled and identified in the field given the lack of any single useable sample frame with good coverage of migrants. The work also included developing, testing and translating a questionnaire for use in a survey of migrants.

The feasibility study was divided into five interdependent modules.

1. Workshops with the UK Border Agency and other stakeholders designed to identify the full range of survey requirements.
2. An exploration of existing definitions of 'migrants' and development of one for use in a survey, through both a literature review and consultation with stakeholders.
3. The development of a sample design to enable the survey to provide good estimates for the population of migrants thus defined and for key sub-groups of that population.
4. Testing fieldwork recruitment and data collection methods to be used in conjunction with the proposed sampling design; this element of the work also involved active participation in a workshop with methodological and statistical experts in order to obtain expert peer feedback on proposed methodologies.
5. The development of a questionnaire, including cognitive testing in several languages.

Please note, the geographical remit of the feasibility study covers England and Scotland only.

The work was undertaken between January 2009 and April 2010. The report which follows focuses on the sampling and field methods aspects of the feasibility study. The report details:

- survey delivery: and key outcomes from migrant workshops;
- defining the survey population of migrants; and
- sampling and field methods.

A separate technical report provides detail on:

- local area consultation; and
- cognitive testing and translation interim reports.

The technical report and the English version of the questionnaire are both published separately.

2. Survey delivery: key outcomes from the migrant workshops

To inform the development of the survey methodology and the survey questionnaire for a survey of migrants, one discussion group and seventeen in-depth interviews were undertaken with:

- representatives of migrant communities from a range of countries within and outside the EU; and
- individuals working with refugees and asylum seekers in England and Scotland.

The discussions raised a number of issues that would be important when designing the methodology and the questionnaire. A small number of key issues are presented below.

Terminology

The official definitions of a migrant and concepts of immigration status may not match migrants' own perceptions of their status. Overall, this suggests a need to capture both migrants' actual status (e.g. through passport nationality) and their perceived status. It also makes clear that whatever the survey inclusion criteria used, it will be necessary to ask questions about migration status using terminology that is understood by migrants themselves.

While those who had come to the UK to work or study all recognised that technically they were considered migrants, they would not spontaneously use the word to describe themselves, preferring the term 'visitors' or 'workers'. For those from North America and Australasia this was because they felt migrants came from poorer countries, while people from the EU felt freedom of movement meant they were not migrants but just workers. There was also a feeling that 'migrant' was perceived as a negative term and therefore they did not want to be associated with it.

When asked how to define the term 'migrant', participants were inclined to focus on motivation for coming to the UK as the first characteristic. It was, therefore, suggested that purpose of visit might be used as an identifier in the definition established for this research.

Terminology around migration was seen to be a contentious area and one that will require sensitive handling in any survey. All those consulted felt 'immigration status' was a term which provoked anxiety and distrust amongst migrants. It was felt to be an accusing term and if used in the survey could lead to migrants being less willing to participate or less open in their answers. As a result, it was recommended that, if possible, the term be avoided or, if a substitute term is not possible, questions on immigration status be included at the end of the survey in order to lessen any negative impact. However, the cognitive testing element of the questionnaire found people were happy to answer questions about their immigration patterns

up front, although in practice the questionnaire that was developed as part of this feasibility study avoided using terms such as 'immigration status'. (The questionnaire is published in the separate technical report).

For those working with refugees and asylum seekers, many use the term 'refugee' to refer to all of their clients on the basis that all their clients are seeking refuge from some form of trauma/persecution. There was also a feeling that 'refugee' is a more 'comfortable term' compared with asylum seeker, as the latter term has developed negative connotations in the public psyche, according to participants although differences between the two terms were recognised.

Key topics for the survey

The exercise to identify the topics that should be covered by a survey revealed a high level of consensus between priorities identified amongst government stakeholders and priorities of migrant communities. Reasons for coming to the UK and factors encouraging people to stay or leave were deemed important, as well as issues around integration, education and awareness of eligibility for services. Participants felt that not enough was known about these issues and a survey could help to fill crucial gaps in knowledge around migration. They were particularly interested in issues where migrants might be missing out, such as the gap between skills and employment and ability to access education services for dependants. In addition they were very interested in experiences of the immigration or asylum system.

Delivering the survey

There was support for the idea of conducting a survey of migrants, although participants felt there were also practical concerns to consider. These included where to conduct the interviews, with several participants raising concerns about the willingness of migrants to be interviewed in their homes. This unwillingness could be from concern about the appearance of their homes – due to overcrowding or poor housing – as much as about inviting a stranger into their home.

Translation of the survey was also a concern, with participants raising the issue of competency in English and ensuring the meaning of questions and answers are kept during the translation process.

The key concern for participants was presentation of the survey, as economic migrants, refugees and asylum seekers may all have concerns about answering a survey from the UK Border Agency. It is recommended that careful consideration be given to how the survey is presented to ensure individuals are aware of data protection guarantees, the independent nature of any data collection and that it will not impact on their right to remain in the UK.

During the cognitive testing and field testing this research was introduced successfully as a 'study of people born outside the UK' (rather than a study of 'migrants') by a number of government departments, including the UK Border Agency.

In the case of refugees and asylum seekers it was felt that to undertake a survey it would be necessary to work in partnership with local refugee charities and agencies and that a cold-calling door-to-door approach could have adverse consequences for survey response. Partnership would only be possible, however, if the survey agency were absolutely clear and upfront about the objectives of the survey and how the findings would to be used.

Other factors mentioned by participants in the refugee and asylum seeker discussions which could enhance participation were gender matching of interpreters with respondents, clear explanation, use of interpreters, use of incentives, the need to stress independence as well as the need to clearly communicate the fact that responses will not affect personal status.

A key finding that emerged from the consultation around refugees and asylum seekers was the need to engage with local community groups that work with these audiences. All participants representing these groups felt that communication with community support organisations would both raise awareness of the survey and encourage participation.

3. Defining the survey population of migrants

Before any sample of migrants can be designed it is necessary to define the survey population, the population about which one wishes to make statistical generalisations. The survey population should be defined in terms of clear eligibility criteria which would enable any candidate individual to be unequivocally defined as being either a member or not a member of the survey population.

In developing a definition for this study, the authors were primarily guided by the expressed wishes of stakeholders about the populations and sub-populations about which they wished to generalise. In particular, this work was informed by internal consultation with the UK Border Agency, a cross-Government stakeholder workshop, consultation with a small sample of migrant community members and also consultation with a number of key academics and with the Office for National Statistics (ONS). However, the authors also gave consideration to definitions already in use by the Government and researchers, to the types of information that is held on potential sample frames, and to the implications for fieldwork of adopting any particular definition.

After considering a number of possible alternative definitions, for the purposes of the feasibility study the population definition was agreed to be:

'A migrant is defined as someone who arrived in the UK in 1990 or later, who was a non-UK national on entry, whose usual place of residence prior to entry was not in the UK, and who has lived in the country for at least three months (one month if an asylum seeker or refugee).'

It should be noted that this definition takes no account of legal immigration status, it includes three groups which stakeholders did not wish to include in the survey: long-term holiday makers, failed asylum seekers and illegal immigrants. However, in practice, applying the definition is unlikely to result in the survey delivering large numbers of ineligible survey respondents and of course it will be possible to identify these in the interview and exclude them after the event. Overall, this definition had considerable advantage over other definitions in that it was relatively simple to apply on the doorstep during fieldwork compared to some of the others considered, and as such, would probably result in more accurate identification and higher survey response rates.

Some key elements of the rationale lying behind the definition are as follows.

- The criterion of *nationality upon entry* to the UK was adopted instead of current immigration status because some individuals in whom one might be interested will have now acquired UK citizenship.

- The term *non-UK national* refers to all individuals who are nationals of another country irrespective of whether or not they are also UK nationals at the same time. This allows the inclusion of dual nationality in the definition and questions on dual nationality can be captured in the questionnaire.
- *Nationality and usual place of residence* are incorporated into the definition rather than a 'foreign-born' criterion as the steering group was less interested in those individuals who were born outside the UK, but may have lived in the UK for all of their lives (children of British army personnel, for example). Country of birth was still captured in the questionnaire.
- *Three months* was selected as the lower limit for economic migrants to avoid including migrants who are just short-term visitors and not here for longer-term work, study or joining family. For asylum seekers this is amended to one month, because some asylum cases may be dealt with in less than three months and after one month they are more likely to have a stable address for contact purposes.
- The choice of upper limit (*the year 1990*) was selected to exclude migrants who had been in the UK for a very long time (e.g. post World War Two) to ensure that the feasibility study focused on more recent migrants who were of more immediate policy interest. As the rate of migration increased in the early 1990s, the year 1990 seemed a logical date to start from.

4. Sampling and field methods

This chapter reviews the sources of information available for designing a survey to fit with the agreed definition. It explores the sampling approaches available to improve fieldwork efficiency and includes results of some field testing. Finally, this information is pulled together in a proposal for a sample design.

In carrying out this work the authors had, in the main, to start from first principles, as there was not any recent and directly relevant work to build on. Although there has been considerable recent work relating to how migration statistics might be improved (see for example Office for National Statistics, 2006), this does not directly address methodological issues relating to a survey of migrants.

4.1 Investigation of potential sampling frames

The agreed definition for the feasibility study concentrated on migration into the UK post-1990. This meant there was a requirement to pick up both established and recent migrants (potentially short-term) as well as recognise the new migration flows of the EU expansion post-2004. The migrant definition implicitly included those claiming asylum, although these individuals may not be represented in the general data sources, while those granted refugee status were naturally included under the main definition.

The role of the sample frame

In any survey the sample frame forms the basis of the sample design and selection of the units. Information available on the frame can enhance the efficiency of a design by helping control selection and thereby reduce the sampling errors. At the very least, the frame needs to provide access to the units of the population, in this case migrants, as per the agreed definition.

Frames for 'direct' sampling

A sample frame is often regarded conceptually as a list of all the units (migrants) in the population that one wishes to represent with a survey. A sample from the population can then be drawn directly from the frame. To achieve this, the list being used as a frame must have very high (near perfect) coverage of the population (in this case migrants as per the specified definition). Alternatively it must have the same properties for a well-defined sub-group of the population so it can be combined with other lists to create a frame for the whole population.¹

1 These requirements can be relaxed further where there are multiple frames, each with poorer coverage but by sampling from all the frames it is possible to get close to the required coverage of the population.

The frame often contains additional information on the members of the population and this can be important in the design of the sample if analysis of specific sub-groups (A8/A2 migrants for example) is required.

Once the sample has been selected, the frame needs to contain relevant contact information for the sample members. This is crucial as without it one cannot contact potential respondents to collect the required information. If the contact information is out of date it creates a risk of significant non-response. In addition, it would not usually be possible to distinguish between those that were still in the population but at a different location (non-response) and those that had left the country (out-of-scope), so judging the true extent of the non-response would be difficult. Investigation of potential frames for a survey of migrants has shown that direct sampling will only be possible for asylum seekers.

Direct frame for asylum seekers

Those seeking asylum form an important sub-group within the main survey and are covered by specific administrative systems within the UK Border Agency. These systems specifically focus on those currently applying for asylum, and hold a range of data relevant to that application. Once granted, those individuals become part of the main immigration system and it is much harder to identify accurate address details as they do not need to keep in contact with the UK Border Agency unless their immigration status is reviewed. But while they are seeking asylum there is generally close interaction between the UK Border Agency and the claimant, which means address details are usually up to date.

Address information is generally current up to shortly after the conclusion of the case with individuals usually in accommodation supplied through the UK Border Agency, making it possible to directly sample this sub-population of migrants.

A small-scale field test was undertaken to assess the practicability of sampling asylum seekers direct from UK Border Agency lists. This work (Appendix A1) concluded that with appropriate fieldwork procedures, direct sampling of this group would prove feasible as part of a survey of migrants.

Frames for 'area' sampling

None of the frames considered for sampling non-asylum-seeker migrants (i.e. economic migrants, students, family joiners etc) provide adequate contact information to enable these migrants to be sampled directly. This is typical of many real sampling applications, where frames of the actual units one wishes to represent do not exist but there are frames of areas that contain the units. For example, when doing social surveys, it is common to have a frame of address locations (the Postcode Address File or PAF) and know that within those address locations researchers will find households and individuals to sample.

Many of the approaches for finding rare populations work on the principle that while it is not known exactly where individuals are to be found (no frame for direct sampling), it is possible

to identify areas where the population is expected to be present. In this case, the frame needs to provide information at some reasonably small level of geographic aggregation (wards or possibly smaller) that identifies the presence of migrants in the area. As the feasibility study aimed to ensure the sample captured specific sub-groups (short-term, citizenship applicant, A8 and A2, those applying under the Points-Based-System (PBS)), ideally the frame needed to be able to classify the areas by 'type of migrant found within' so that the areas within the sample design could be stratified.

At the estimation stage, a good area frame may provide high quality counts of the totals of certain sub-groups of migrants within the country (even if at the local level the specific location of an individual is poor thereby negating the possibility of direct sampling). If this is true, it will provide information allowing for calibration approaches to estimation, which in turn generally leads to improvements in the quality of the estimates produced.

When creating an area frame for sampling rare populations there is the possibility of both false positives and false negatives. False positives are where an area is classified as containing the rare population when in fact it does not. This inflates the costs of the survey and damages the overall efficiency of the sample with interviewers searching in an area for individuals that do not exist within that area. False negatives are where an area is classified as unimportant with respect to the rare population when, in fact, it is an area with large numbers of the population within it. This damages the coverage of the survey and exposes the estimates to potential bias if specific sub-groups are missed by this occurring. To minimise both, the frame needs to be based on the most up-to-date information available.

Local Authority profiles

The ONS is introducing a suite of statistical indicators at the Local Authority level. These consist of indicators using parts of the data sources discussed below. In addition, estimates of the 'stock' of non-UK born and non-British adults based on the Annual Population Survey are available. These can essentially be seen as updates of the 2001 Census figures but at a high level of aggregation (and with sampling error that may be reduced by averaging adjacent years). The level of aggregation is too high to be of direct use in any sample design but it does give another source that can be used as a comparison for data that are available at lower levels of aggregation. Further information on this project (as well as access to the current data) can be found at <http://www.statistics.gov.uk/statbase/product.asp?vlnk=15239>.

Combining the data – creating an area frame

The following sources were considered for use in the area frame:

- Home Office administrative data;
- Worker Registration Scheme (WRS);
- National Insurance Numbers (NINO's);

- Flag 4 (General Practitioner (GP) registrations data);
- Pupil Census Data (PLASC);
- Electoral Register (ER);
- 2001 UK Census;
- Labour Force Survey (LFS);
- Annual Population Survey (APS);
- International Passenger Survey (IPS);
- Higher Education Statistics Authority (HESA) Database; and
- Citizenship Survey (CS).

All the sources considered have strengths and weaknesses with respect to use for an area frame and, therefore this section reviews how the authors combined information from some of these sources to produce a proxy size of the migrant population consistent with the proposed definition at a ward level. To do this, techniques for small area population estimates are used to produce updated 'sizes'² for wards based on the current APS data at the Local Authority level and other indicators at the ward level.³ One approach apportioned out the current total for the LA based on the distribution of a current indicator. For example, one may share out the current APS total based on the 2001 Census distribution assuming that new migrants are going to the same areas in the same relative numbers as in the past. It is then possible to build on this, but combine information from several sources through a regression model with no intercept term. The model can be fitted at the LA level, where there are very strong inter-relationships between the various sources, and then the model is used to predict a proxy for the APS figures at a ward level.

The first stage of the process is to set up the regression model at the Local Authority level. This will define a set of parameters that link the patterns on the administrative data at this level with the patterns seen in the APS estimates. So far, small-area estimation has not been conducted but a synthetic assumption is made (this is un-testable from the data) that the fitted patterns at the Local Authority level can be applied at ward level to get 'predictions' of migrants at that level. Therefore, while the inter-relationship between a particular administrative data source and migrants on the ground is constant across the data and defined by the modelling (see Table 4.1), the actual prediction at the ward level will depend on that ward's specific profile with respect to the various sources.

Data used to fit the Local Authority model

- The APS count estimates (Y) of those 'not born in the UK' averaged across 2006, 2007, and 2008 but restricted to those individuals that have arrived since 1990 as this more closely reflects the definition specified within the feasibility study
- The 2001 Census count (X_1) of those 'not born in the UK' to give a historical stock count.

2 Note that the authors do not consider these as estimates of the migrant population but perhaps useful indicators of the wards with the larger populations as well as an indication of relative sizes.

3 A short review of these approaches by James Brown and Ian Diamond can be found in '*Population counts in small areas: implications for studies of environment and health*' published by The Stationery Office (1999).

- The NINo data (X_2) give a count of new applications allocated to the Local Authority from different parts of the world, with a focus on the counts for the EU. These counts are averaged across 2006, 2007, and 2008 split into whether the individual was from an established EU15 country or a new EU Accession (A8/A2) country. The two resulting variables give the impact of recent migration from the EU, and in particular allows for the influence of the new EU migration.
- Data from the UK Border Agency (X_3) give the count of individuals within each Local Authority granted 'leave to remain' per year averaged over 2006, 2007, and 2008 to pick-up the more recent arrivals of 3rd country nationals.
- Data from the UK Border Agency (X_4) gives the count of individuals within each Local Authority granted 'indefinite leave to remain' (ILR) per year. This is added to the count of those granted Citizenship and then averaged over 2006, 2007, and 2008 to pick-up the longer-term 3rd country nationals.
- Interactions (X_5) to allow for a differing relationship within Scotland. This gives the potential for Scotland to end up with a completely different model but without the statistical inefficiency of splitting the data and assuming it will be different.

Data NOT used to fit the LA model

A number of other potential sources which could be useful for identifying migrant areas were considered but rejected.

- 'Flag 4' data⁴ could be included in the LA model, however their inclusion substantially reduces the importance of the NINo data, and there were concerns over their completeness. This causes two problems:
 - (i) 'Flag 4' data do not allow the new EU migration to be explicitly represented and
 - (ii) they force a separate model to be fitted for Scotland for definitional reasons. However, with 'Flag 4' data available at the ward level they can still be used to see if they give additional confidence in the model's estimates for recent migration patterns.
- The Worker Registration Scheme (WRS)⁵ gives information on workers from the new EU. This does seem to be correlated with the APS data at the LA level and adds to the model, but there are problems with using this variable for ward-level predictions, as it is associated with employer rather than employee addresses.
- The Pupil Census data for pupils are only available for England, making their inclusion in a combined model problematic, but as with 'Flag 4' they can be used to compare to the ward patterns of the proxy produced by the modelling.

4 'Flag 4' refers to new National Health Service (NHS) registrations (registering with a GP) from those who previously lived outside of the UK.

5 A8 migrants who wish to work in the UK are generally required to register with the WRS.

- The Higher Education Statistics Authority (HESA) data is useful in identifying local areas that have a high count purely due to the presence of international students, rather than a general predictor of local migration patterns. It would then be possible to exclude areas that were purely driven by students, if these are not important to the survey, or at least for the interviewers to be better prepared in accessing student halls before going to these areas.

Steps involved in making the ward level predictions

1. Fit the following linear regression model to the data at the Local Authority level:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \text{error}$$
excluding any LAs with a zero APS estimate in all three years used in the average of the APS estimates to define Y. A simple forwards step-wise approach is used to add the various data sources and explore the need for Scotland interactions.

As this is a linear regression model through the origin it is not sensible to assume a constant residual variance, which is what the standard OLS regression model does. Breaking this assumption does not impact on the estimates of the model parameters resulting from using OLS (apart from a marginal loss of efficiency), but does mean the p-values on the model parameters should be treated with caution. However, in this case the main focus was on a plausible model for prediction and all the observed p-values were well below 0.05, so a little efficiency may be lost if the final terms included in the step-wise modelling would be only marginally significant using corrected standard errors.

2. Define a scaling factor⁶ that calibrates the LA predictions from the fitted model to the observed LA values. These factors just ensure the fitted values recover the observed values and the use of such factors is standard within small area estimation. In this case the factors are generally close to one as the R-square and adjusted R-square in the final regression exceed 98 per cent. The high R-square value also demonstrates, at the Local Authority level, the strong relationship between the APS estimates of migration patterns and the administrative data sources used to capture those patterns.
3. Construct the administrative data sources using the counts and averaging over the three years,⁷ where appropriate, at the ward level. The ward 'counts' for any source will sum to the Local Authority 'count' for the same source.
4. Use the model parameters from the Local Authority model applied to the ward level data from the administrative sources to 'predict' the ward counts based on the ward level distribution of the X's (2001 Census, NINo's, UK Border Agency data) and apply the scaling factor to ensure the predicted values for the wards sum to the observed value for the Local Authority.

6 This factor is set equal to one for those LAs that were excluded from the model fitting due to three estimates of zero in the APS as the model can still predict for them given non-zero values for at least one of the X data sources.

7 With the NINo data, the authors only had ward level counts for a single year and the wards do not match perfectly, so this contribution is not as stable as one would like in a final analysis and it does not sum to the Local Authority values used to fit the model.

Discussion of the fitted LA model

Table 4.1 *Modelling the relationship between administrative sources and the APS at Local Authority level*

Model	B	Std. Error	t	Sig.
1 Census (non-UK born)	0.749	0.009	85.615	0.000
2 Census (non-UK born)	0.484	0.017	28.336	0.000
Leave to Remain	3.965	0.238	16.675	0.000
3 Census (non-UK born)	0.377	0.029	12.908	0.000
Leave to Remain	3.340	0.270	12.360	0.000
Indefinite LR & Citizenship	3.210	0.720	4.455	0.000
4 Census (non-UK born)	0.247	0.032	7.741	0.000
Leave to Remain	1.691	0.331	5.103	0.000
Indefinite LR & Citizenship	6.246	0.777	8.043	0.000
New NINOs (EU15)	4.493	0.595	7.550	0.000
5 Census (non-UK born)	0.223	0.032	6.916	0.000
Leave to Remain	1.518	0.330	4.606	0.000
Indefinite LR & Citizenship	6.175	0.764	8.087	0.000
New NINOs (EU15)	4.801	0.592	8.114	0.000
New NINOs (A8/A2)	0.731	0.210	3.472	0.001
6 Census (non-UK born)	0.229	0.032	7.134	0.000
Leave to Remain	1.503	0.327	4.588	0.000
Indefinite LR & Citizenship	6.345	0.762	8.322	0.000
New NINOs (EU15)	4.548	0.598	7.600	0.000
New NINOs (A8/A2)	0.522	0.229	2.285	0.023
Indefinite LR & Citizenship (Scotland Interaction)	3.030	1.338	2.265	0.024

- Dependent variable: APS average over three years (post-1990 entrants).
- Linear regression through the origin.
- The A8 and A2 countries are combined to pick up the overall impact of EU expansion, although it is recognised that at a local level this will likely be migrants from a single country rather than a mixture across all the countries.

Table 4.1 shows the building of the LA model using a forward step-wise approach finishing with the model that can be used for the ward level predictions. From the model parameters one can see that the current migration stock comes from the Census, and this is the first variable in the step-wise, but with particular impacts from the long-term UK Border Agency data and the EU15 data. The smaller parameter on the new EU presumably reflects the more transient nature of these migrations with many staying for short periods and therefore making a smaller contribution to the current stock as measured by the APS. However, this seems plausible given what is known about the behaviour of this group. The interaction term added in the final model shows the increased importance of the long-term UK Border Agency data in capturing the patterns across the Local Authorities in Scotland.

This structure makes identifying the type of migrant coming from a ward more straightforward. The Census gives the very long-term estimate for all migrant types, the UK Border Agency provides the new and more settled 3rd Country flows since the Census, the NINo data supply the more recent EU flows from both the old and new EU countries. As noted above, the authors are using the NINo data in preference to WRS as it more directly relates to where the migrant resides, and in preference to 'Flag 4' to get more information on the area of the world from which the migrants are coming.

Concluding remarks on frames

It is difficult to assess whether the identified ward level patterns look sensible as no definitive data exist to compare this with. As already mentioned, the education data and Flag 4 data have not been used and therefore present an opportunity to sense check patterns, although each has their weaknesses. In putting together the area frame for the actual survey the authors recommend sticking to the sources used here for the modelling, and then checking patterns against PLASC data and 'Flag 4' data. This has the potential for highlighting anomalies for further investigation. The HESA data could play a role in identifying high concentrations of students, and would provide an opportunity to exclude such areas if they were not a sub-population of interest. Both PLASC and HESA are not suitable for inclusion in the model as they identify schools or universities attended rather than migrant households. PLASC data were also excluded from the model because they provide only proxy data for migrants – counting English as an additional language in combination with detailed ethnic group, rather than migration status per se.

A further project involved consulting local experts and these individuals were given the opportunity to comment on the local patterns coming from the modelling (a review of this project is outlined in the technical report). Broadly speaking, these 'experts' were supportive of the patterns found by the modelling or did not have sufficient local knowledge to challenge them. This added credibility to the modelling and therefore was useful for the feasibility study, but the authors would not recommend it as a general part of a main survey. To be done across all areas in a larger survey, all Local Authorities would need to provide the same quality of information to avoid biasing the patterns in the frame. However, if comparisons with the additional sources not used in the modelling do highlight a small number of local anomalies, the authors' experience here suggests that local experts may be able to explain the reason and therefore allow decisions to be made on the appropriateness of the models for what we would expect to be a small number of wards.

The authors see this approach of combining data as the best way to get the most from the available sources and provide an idea of ward level patterns. Further work would be needed to secure more NINo data across years for use in a final sample frame but the predictions from this model were considered fit for use as a basis for the proposed sample design in the sections that follow.

4.2 Possible sampling methods for a survey of migrants

When sampling rare populations, the approach to sampling will be driven by the frame information. Possible frames were discussed in section 4.1 and distinguished between frames which identify the rare population members directly and ones which can highlight areas where the population is concentrated. In the former, it is possible to use standard approaches to select a sample 'directly' from the rare population and the authors conclude that this is feasible for the asylum sub-group but not for other migrant types. In this case the sampling approach is straightforward as the contact details for potential respondents are available directly from the administrative records. With frames of the latter type, ones which can be used to identify areas but not for direct sampling, selecting areas can be fairly straightforward and made more (or less) efficient by the information on the frame.

In this section there is a focus on the impact of using frames of the second type as direct sampling will not be possible for most migrant groups. As the authors concluded in the previous section, it is possible to construct an area-based frame by combining data sources to estimate concentrations of the target population, and then within areas identified by this approach it will be necessary to use screening methods in the field to generate the sample of the rare population. The effectiveness of these methods will depend on different factors 'on the ground' in relation to the localised distribution of the rare population within the selected areas.

In this section various approaches are reviewed that exist in the literature for selecting a sample from a rare population within areas. A brief discussion of the approaches and their usefulness with respect to the issue of a survey of migrants is detailed in Appendices A2 and A3 which include information from some small-scale field tests. A suitable approach is then outlined, drawing on the review and the field experience. Kalton and Anderson (1986) and Kalton (2009) give reviews of many of these approaches.

Random probability or quota sampling within areas

Within 'Official Statistics' there is general acceptance that random probability methods provide a robust framework for inference regarding a population (Smith, 1983), although quota sampling methods can, in certain circumstances, provide valid inference. Smith (1983) demonstrates that this choice hinges on whether the sample selection is 'ignorable' and in quota sampling this implies that within the groups set by the quota variables it does not matter which individuals are sampled. If this is true, one can select individuals in the most convenient way and with quota approaches this will often involve random walks by interviewers around an area recruiting respondents to meet quota numbers.

Extensive discussion throughout the study has resulted in the rejection of a quota sampling approach as being inappropriate in this case. Firstly, while the estimation of population totals is not required, robust estimation of proportions within the migrant population with appropriate confidence intervals is desirable and this would not be possible with a quota approach because there is insufficient information to set quotas in local areas for

interviewers to fill. Therefore, any structure in the sample is un-linkable to the population behind the sample. Therefore, even if it was felt the sample being selected within the set quotas was valid for inference within that quota group, it is not possible to then combine the quotas to produce estimates relating to the population.

A random probability approach will mean less control over the sample composition in terms of types of migrants or nationalities as these will tend to fall out naturally as they occur in the migrant population. A quota approach, on the other hand, could be used to target certain groups, albeit in a way which would not allow generalisation to the population. It might be possible to target certain groups with a random probability approach through selection of sampling points with high concentration of the desirable types of migrants, to the extent that this information is available, but this could not guarantee sufficient sample sizes within groups. It could also increase costs if the desirable groups are in less concentrated areas.

Data collection methods for screening

Given that one would need to sample areas and then attempt to identify the migrant population, screening will be necessary to identify the migrants. This generally involves a large-scale sample with a simple questionnaire to decide on whether a respondent qualifies for membership of the target population, tested in field tests during the feasibility study using face-to-face interviewing. Given that the definition of a migrant being proposed is relatively simple to apply in the field, this screen could potentially be done by post or telephone. Both are relatively cheap options (and would make the approach to the sampling very different) but have clear problems with either coverage or response (or both). Therefore, these were not pursued as options during the feasibility study.

In the early stages, the UK Border Agency considered the possibility of using a postal and/or telephone approach for the main survey. For example, once a respondent is identified in the field he/she could then be given a questionnaire to complete and return in the post. However, given the complexity of the questionnaire that had been developed, such an approach was not considered to be viable. The whole point of a specific survey of migrants is to collect the kind of detailed information that would not be possible without a face-to-face interview. To identify the experiences of different groups will potentially require complex routings through the questionnaire so even without the additional complexities of language a self-completion approach is not advisable.

Screening using existing surveys

Screening requires a very high response rate to be effective and this is why there is generally a preference for using interviewers and face-to-face contact. This immediately makes the approach expensive unless there is a high degree of geographic clustering in the screening survey. However, there are already a number of large-scale surveys that identify migrants⁸ and, as such, might potentially operate as a screening survey. Two examples of such surveys would be the Labour Force Survey (LFS) and the Citizenship Survey (CS). The issue of data

8 The LFS now collects an expanded set of questions that would allow one to identify migrants based on the agreed definition.

access, as well as whether individuals could be re-contacted, would need to be agreed. The data access might be less of an issue if the migrant follow-up became an integrated sub-sample within the main sample, or if respondents were asked in the survey whether they minded being contacted for further research.

Such an approach could be very cost-effective and certainly give good coverage of the household population. However, in the migrant population those within communal establishments (CEs) are potentially important, although this may again depend on the types of migrants within the scope of the survey. For example, CEs would be very important if one's definition of a migrant included recent arrivals, particularly asylum seekers, but much less important if one was only considering those who had been in the country for at least a year, say. The LFS does cover some of the CEs but the exact nature of the coverage would need to be considered. See http://www.statistics.gov.uk/downloads/theme_population/LFS.pdf for some discussion of the LFS coverage of migrants and the particular problem of its poor coverage of CEs.

Such an approach potentially would generate a large pool of migrants over the year from which to select a migrant sample. As well as collecting the information to identify migrants, the LFS also generates additional covariate information that would potentially allow direct control of the sample in relation to other variables such as region and working status. It was not possible to pursue this further during the feasibility study, but recent contact with ONS suggests this may be a genuine possibility. One disadvantage would be timing of the waves of the LFS, which could limit coverage of short-term migrants if they were only selected at the completion of the five LFS quarters. In addition, over the five LFS fieldwork waves there is respondent drop-out introducing potential bias into the pool of migrants available to sample from the end of the LFS fieldwork. However, these disadvantages would need to be weighed against the high response rates within the LFS relative to other surveys, the potential cost-savings, the ability to control directly the selection of migrants, and the fact that information on respondents from earlier waves of the survey will allow some assessment of any bias.

In terms of the Citizenship Survey, whilst it is important as a key comparator survey on integration and cohesion issues, it is less useful as a screening survey as the sample size is smaller than the LFS.

Special screening samples

More usual is the selection of a screening sample specifically for the purpose of identifying members of the migrant population and this has been the focus of the feasibility study. The 'standard' approach to screening involves issuing fixed samples of addresses to interviewers and having them visit each address in order to establish the eligibility of the occupants. This approach has often been used with success in the UK (e.g. in the Citizenship Survey) and is an effective method to use in areas where the target population is reasonably highly concentrated (say, where 10% or more of addresses contain eligible people).

With screening samples, the frame is generally used to identify 'areas' where the migrants are likely to be found but in which eligible individuals/households cannot directly be identified.

This information allows the screening sample to be heavily concentrated into areas most likely to yield migrants. It also allows for areas with very few members of the population to be excluded to improve efficiency for very little loss of coverage (as defined by the information on the frames). This is essentially the approach taken in the design of the ethnic minority boost sample for the Citizenship Survey. Using smaller units than the traditional wards, such as Output Areas, may also help further in terms of field efficiency but does potentially increase the clustering effects in the final sample.

Given the proven track record of conventional face-to-face screening methods in areas of higher concentration, it was concluded that this method was appropriate for a survey of migrants. However, this approach becomes expensive in areas of low concentration, and therefore it was also considered whether alternative screening methods might be more cost-effective in such areas.

Strategies for focusing the screening

One method which has been used regularly to screen for the presence of ethnic minorities in areas where they are thinly spread is focused enumeration. This entails asking households whether their neighbouring households contain members of the eligible population. The approach was originally developed in the 1980s (Brown and Ritchie, 1981) and has been further developed and used extensively in surveys to identify members of ethnic minority populations relatively cheaply. However, the problem with using this approach to identify migrants is that there is a lack of obvious physical characteristics (like skin colour) which can be used to distinguish many migrant groups (e.g. most A8/A2 migrants) from their neighbours. This is, therefore, unlikely to be effective or efficient as a method for screening in areas of low concentration.

The starting points for the authors' investigations were the reviews of methods for sampling rare populations by Kalton & Anderson (1986) and Kalton (2009). Here they focused on two types of approach to improving the screening efficiency: methods that utilise the social networks of respondents; and methods that utilise localised clustering of the rare population. Two approaches looked to be most promising – multiplicity sampling (social networks) and adaptive cluster sampling (local clustering). Field tests were conducted to evaluate these methods, which concluded that the multiplicity approach would not deliver sufficient additional migrants to make it worth implementing in the field, while adaptive cluster sampling potentially could. Hence, this method is considered further here while details of both approaches are given in Appendices A2 and A3.

Adaptive cluster sampling – method and results

This is an area-based approach that has been used to count wildlife populations. It has a sound theoretical basis as a random probability approach and, as such, can give unbiased estimates of a rare population. In the current context the key steps involved in adaptive cluster sampling are outlined below.

- Divide all addresses within each pre-specified sample area into equal-sized blocks (e.g. of five addresses).

- Select a sample of blocks within each sample area.
- Screen all addresses in each sampled block for migrants.
- If a particular block reveals a migrant, move to the two blocks adjoining it (one each side on the PAF list) and keep repeating until blocks containing no migrants are screened.
- If the initial block contains no migrants move on.

To evaluate whether this could work in the field, field tests were conducted in four wards in England and Scotland which were expected to contain migrants (using UK Border Agency activity data and Census data on the non-UK-born population). In the initial test in Gateshead, 20 blocks of 10 addresses were sampled from the Postcode Address File in a ward; for the remaining tests this was adapted to ten blocks of five addresses within half-wards. The table below summarises the fieldwork outcomes for all addresses screened – both initially sampled addresses and addresses identified in the blocks subsequently brought into the sample.

Table 4.2 Adaptive cluster sampling field test outcomes

	Gateshead	Coventry	Dundee	Edinburgh
Initially issued addresses	200	50	50	50
Additional addresses generated in new blocks	210	70	0	45
Deadwood addresses	18	9	4	2
Non-deadwood addresses	392	111	46	93
Number of addresses successfully screened	217	86	33	81
Screening not attempted (block eligibility already established)	38			
<i>Screening response rate</i>	62%	77%	72%	87%
Total No. of eligible households identified in screening	29	12	0	7
<i>Proportion of those screened identified as eligible</i>	13.4%	14.0%	0.0%	8.6%
Achieved interviews	17	11	0	7
<i>Interview response rate</i>	59%	92%	N/A	100%

In Dundee, no eligible migrants were identified in the initially sampled blocks, and as a result no new blocks were brought into the sample. The area of Dundee selected contained both university halls of residence and a medical training facility with a campus. Anecdotal accounts suggested that the selected ward *did* include many migrants, but that these individuals lived in properties which were excluded from the sample frame or were inaccessible to the field tests (e.g. in halls of residence or medical institutions).

Adaptive cluster sampling: conclusions

The adaptive cluster sampling method was readily implemented by interviewers and it was concluded that it could be operationalised in a larger survey. The crucial question, however, is whether it was effective at generating migrant interviews. Here the findings suggest that the rate of identification or number of additional interviews achieved at new addresses is positive but not hugely impressive – on average across the tests 13 screened addresses were required per interview from new addresses added via the method, compared with one in every 24⁹ of the initially sampled addresses. This suggests that the method is indeed identifying clusters of migrants, but that such clustering is not very pronounced.

However, a major potential difficulty which would need to be overcome when applying the method in practice would be setting a block size which generated a sufficient, but not excessive, number of interviews. In order to follow random probability sampling principles, it is essential that the method is followed to the end. The point at which this happens depends on two linked factors – the local concentration and distribution of migrants in the areas surrounding the initially

sampled blocks and the choice of block size. If the blocks are too small for a given migrant concentration, then one risks generating too few or no migrants. If the blocks are too large it is likely that they will continue to generate new blocks until the entire ward, or a very large part of it, has been screened which would result in (i) unacceptably large clusters of interviews, with resultant damage to the precision of survey estimates and (ii) a protracted fieldwork period (in a typical ward of around 2,000 or more addresses around 10 to 20 iterations would be required to screen the entire ward). Three of the four test areas were fully worked until no new migrants were identified (in Dundee this was true only in a trivial sense because no migrants were discovered in the initial sample of addresses). The fourth area (Gateshead), which used a larger block size of ten, was shown to generate more new addresses after each round of new block generation than it started with. The area would most likely have been controlled with blocks of five addresses, but the method is open to this risk.

In conclusion, adaptive cluster sampling could be readily implemented in the field but in practice it would be very hard to control both cluster sizes and final sample size. It does offer some hope in the less concentrated areas where a random screening sample spread-out over the whole area would be particularly inefficient.

Furthermore, small-scale testing of focused enumeration was integrated with this field test and fully confirmed that focused enumeration will not boost numbers with the migrant population because, as expected, migrants are not ‘visible’ enough in their local communities.

Using the Waksberg methodology to develop adaptive cluster sampling

Although the Waksberg approach (Waksberg, 1978) was initially developed for use with telephone surveys, it could, in principle, be applied more widely. The method offers a way of selecting a random sample of a rare population by selecting small areas with probability proportional to the size of the rare population within them, and then taking a fixed sample within each area.

9 This figure is inflated somewhat by the 46 non-deadwood addresses in Dundee which yielded no interviews, leaving Dundee out of the analysis; the interview yield out of initial addresses is 1 in 20.

As applied to a face-to-face survey of a rare population group, the Waksberg method might be implemented as follows.

1. Divide addresses on the PAF sample frame into blocks of c addresses where c might be in the order of 50–100 addresses (it is generally set at 100 in telephone surveys).
2. Select an equal probability sample of these blocks (it would be possible to stratify the blocks to control for specific sub-groups of the migrant population if the frame contained relevant information).
3. Randomly select a single address in each sampled block.
4. Have an interviewer screen the address for the presence of eligible occupants (eligible migrants in this case).
5. If the address proves to be eligible, randomly select further addresses from the same block until a fixed number, k , of *eligible* addresses have been identified in the block.
6. If the initially sampled address in a block does not contain a migrant, no further addresses are sampled in the block.

In the form just described the method delivers an equal probability sample of eligible addresses. It would, however, prove hard to control the second stage of address selection (step 5 above) in a face-to-face interview setting. Because the requirement is that k *eligible* addresses should be identified, the number of selected addresses in each block is not fixed, but rather depends on the outcome of screening at previously covered addresses. And because this screening may take some weeks to implement at some addresses (because of holidays, persistent non-contact, etc.), the method in its pure form would prove very hard to implement. The authors do, however, feel that the method could be profitably combined with adaptive cluster sampling (ACS).

Both the Waksberg approach and adaptive cluster sampling start by creating blocks, in this case of addresses, within selected fieldwork areas (most likely wards based on the frame development discussed in the previous section). The assumption underlying these methods is that within some blocks there will be clusters of migrants. Blocks would be larger with the Waksberg method than with the adaptive cluster sampling method.

As described in the previous section, the lack of control over fieldwork is a concern with adaptive cluster sampling. Also, adjusting for non-response in screening will need further thought as ACS requires a decision to be made as to whether a block contains a migrant or not, even when all addresses cannot be screened due to non-response. However, the Waksberg approach in its pure form is no easier to implement as random screening of an area for a *fixed number of eligibles* is difficult for interviewers to carry out without the exercise effectively becoming one of identifying a quota of eligibles from the block. These difficulties could be largely removed by combining the two methods.

The authors, therefore, recommend that the following variation of the Waksberg approach should be seriously explored for use in a survey of migrants within areas of low migrant concentration:

- select a random sample of blocks within a sampled area;
- select a single address at random to screen within each sampled block;
- if the screened address reveals a migrant, screen the entire block but do not carry on to neighbouring blocks;
- if the screened address is a non-migrant, leave the block.

This essentially starts by picking single addresses spread throughout the ward and then enumerating a block around them if a migrant is found at a sampled address. The blocks within a ward would be pre-defined and the researchers would select blocks and then an address within the block at random. This approach provides more control than does ACS over the number of addresses an interviewer would be screening whilst at the same time taking advantage of the fact that local clustering does appear to exist on the ground (as demonstrated in the field tests).

It is suggested that the blocks should be relatively large (say, 20–50 addresses), bigger than the initial test blocks used by the authors.

In summary, the proposed hybrid method will be easy to apply in the field. It is less complicated than ACS, which worked well, and, as far as the interviewer is concerned, it is very similar to tried and tested face-to-face screening methods.

In order to control further the sample efficiency in terms of weighting, if initially sampled wards are selected with probability proportional to 'expected number of migrants' it is possible to adjust the selection of the blocks to prevent final selection probabilities becoming too variable. This is similar to the approach that would be taken in selecting for standard screening of addresses, as discussed later, but would be based on the selection of a blocks and one address per block.

Screening methods conclusions

It is likely that there will continue to be a place for traditional face-to-face screening methods¹⁰ in areas of higher migrant concentration. However, in areas of lower concentration (a typical population concentration level above which face-to-face screening methods might be used in the UK is 10%) significant advantages may attach to using the hybrid Waksberg/ACS method just discussed in order to improve the efficiency of fieldwork.

Sampling asylum seekers

As mentioned earlier this is more straightforward as it is possible to use direct sampling methods. The sample could be stratified by geographical dispersal area, and perhaps case type, depending on the requirements for the survey. The authors recommend that the UK

¹⁰ In essence these involve drawing a large random probability sample of addresses and having interviewers visit each to establish whether or not its residents are eligible for the survey.

Border Agency's main asylum caseload database, separate Local Authority level data on asylum seekers in supported accommodation and data on unaccompanied asylum seeking children (UASC) – to allow removal of minors from the sample databases – are used for this purpose. The reference numbers contained in these databases should be clarified for matching. It would also be beneficial if further, more comprehensive data which would enable identification of minors, could be obtained.

Small-scale field tests (see Appendix A1 for further details) suggest that around two-thirds of a sample from the caseload database could be valid/ potentially productive, which suggests that a response rate in the region of 40 per cent would be possible, providing the necessary preparatory work is undertaken: engaging agencies and relevant community groups to enable access to asylum seekers; adequate language cover; and putting in place following procedures for those who have moved.

4.3 Towards a sample design

Introduction and general approach

A key challenge to be addressed when designing a sample of migrants relates to identification of eligible sample members. With the exception of asylum seekers, migrants cannot be sampled directly from available lists. This means that for the most part migrants need to be identified through a screening exercise. As discussed above, unless it proves possible to arrange access to migrants identified through a previous survey (e.g. LFS), the identification of migrants will require some form of screening in the field.

Field screening is very expensive (the costs of screening would almost certainly be greater than the costs of conducting the migrant interviews), and it is, therefore, essential that the sample design maximises its efficiency. Hence the feasibility study invested effort in trialling methods which may increase the efficiency of standard screening methods, while at the same time establishing whether sample frames can point to more concentrated areas. While multiplicity sampling and adaptive cluster sampling have been found wanting, the authors recommend that in future pilot work the efficacy of the hybrid method outlined above should be explored. The authors are also confident that there will be a major role for conventional screening methods which have been used in surveys of ethnic minorities.

Bearing in mind the high costs of screening, efforts to maximise screening efficiency will largely have to be focused on designing a sample which concentrates screening efforts in areas where most migrants are to be found. This approach inevitably requires making a compromise: accepting it is only possible to cover the portion of the population of migrants living in areas where they are more concentrated. For example, in England the authors estimate that if they were to cover all migrants, they would need to screen about

13 addresses to identify a single migrant address. If, on the other hand generalisations were limited to the 82 per cent of the migrant population living in areas of higher migrant concentration they would only need to screen about six households in order to identify a migrant address.

Another point to note is that, in common with most government surveys, the sample would typically be issued in geographical clusters. Clustered samples are cheaper to interview than are unclustered ones, and therefore deliver larger samples for a fixed budget. Despite the fact that they are less statistically efficient (i.e. deliver wider confidence intervals for a fixed sample size) than unclustered samples, clustered samples will generally deliver more precise estimates than will smaller unclustered samples of the same cost.

Illustrative sample designs for the survey of migrants

Introduction

There is no ‘ideal’ sample size – choice of sample size needs to take account of survey priorities, other design decisions and available resources. For the feasibility study, one of the requirements was that the survey design should cover England and Scotland, but not Wales and Northern Ireland, and that separate estimates will be required for these two countries. Initial indications suggested that achieved samples of 6,000 in England and 3,000 in Scotland might be affordable. These would deliver a reasonable range of sub-group estimates. Therefore, for the purposes of exposition, the authors have assumed these sample sizes in the discussion that follows.

From the authors’ discussions with the UK Border Agency and other stakeholders it emerged that it might be desirable for a survey to be able to deliver sub-group estimates for the following groups.

- EEA (European Economic Area), ‘new’ (A8/A2 accession countries), EEA ‘old’ (the EU15), non-EEA.
- Immigration status – work, study, family, asylum, refugees.
- British Citizens and non-British Citizens.
- Recent and long-term migrants.
- Government Office Region (GOR).
- Urban and rural.

Towards a design

The shape of the final design will very much depend on the available budget. The authors assumed in the following that this will be sufficient to allow coverage of about 80 per cent of the migrant population in England and Scotland.

The authors also assumed that it will not be feasible to cover migrants in large communal establishments – apart from asylum seekers for whom direct address lists are available – although it should be possible to pick up some of those living in small communal

establishments by means of the recommended sampling procedures.¹¹ The authors say this confidently because inclusion of residents of communal establishments has proved intractable even in straightforward general population samples. Why this is the case has been discussed recently by Pickering, Smith, Bryson and Farmer (2008).

If one makes the assumption that the regression-based estimates of the numbers of migrant households living in individual wards across Great Britain are accurate, one can examine the distribution of migrants across wards of different migrant concentrations. These distributions are shown separately for England and Scotland in the tables below.

Table 4.3 Concentration and coverage of migrant households (England)

Concentration of migrant households	Estimated migrant households	PAF Addresses excluding deadwood (assumed to be 8%)	Migrant 'Coverage'	Number of wards	Cumulative % of wards
50% or more	73,639	123,263	4.74%	37	0.49%
45%, under 50%	31,528	67,262	6.77%	18	0.73%
40%, under 45%	67,327	159,668	11.11%	39	1.24%
35%, under 40%	102,901	275,585	17.74%	64	2.09%
30%, under 35%	156,829	482,856	27.84%	102	3.44%
25%, under 30%	153,843	559,625	37.75%	120	5.03%
20% , under 25%	159,627	706,803	48.03%	156	7.10%
15%, under 20%	139,510	806,915	57.02%	187	9.57%
12.5%, under 15%	76,327	555,312	61.94%	142	11.45%
10%, under 12.5%	93,269	837,697	67.94%	220	14.37%
7.5%, under 10%	99,952	1,156,995	74.38%	329	18.72%
5%, under 7.5%	125,441	2,045,705	82.46%	704	28.05%
2.5%, under 5%	151,711	4,288,754	92.23%	1,619	49.48%
2%, under 2.5%	35,216	1,573,219	94.50%	658	58.20%
1.5%, under 2%	37,494	2,156,137	96.92%	891	69.99%
1%, under 1.5%	31,262	2,508,543	98.93%	1,115	84.76%
0.5%, under 1%	14,785	1,882,980	99.88%	899	96.66%
under 0.5%	1,813	468,925	100.00%	252	100.00%
Total	1,552,474	20,656,241		7,552	

11 Large communal establishments, such as student halls of residence, do not appear on the standard small user PAF used for sampling. Inclusion of individuals living in large communal establishments requires a completely separate sampling approach.

Concentration of migrant households	Estimated migrant households	PAF Addresses excluding deadwood (assumed to be 8%)	Migrant 'Coverage'	Number of wards	Cum % of wards
15%, under 20%	4,738	28,638	6.61%	2	0.57%
12.5%, under 15%	8,850	65,643	18.96%	6	2.27%
10%, under 12.5%	5,781	50,387	27.03%	4	3.40%
7.5%, under 10%	4,908	56,005	33.88%	5	4.82%
5%, under 7.5%	13,800	229,100	53.14%	22	11.05%
2.5%, under 5%	13,469	381,817	71.93%	55	26.63%
2%, under 2.5%	5,198	229,419	79.19%	39	37.68%
1.5%, under 2%	5,053	297,832	86.24%	51	52.12%
1%, under 1.5%	6,900	547,525	95.87%	95	79.04%
0.5%, under 1%	2,850	358,556	99.84%	68	98.30%
under 0.5%	112	24,810	100.00%	6	100.00%
Total	71,659	2,269,732		353	

Of particular interest in these two tables are the migrant coverage figures in the fourth column. These figures show what proportion of the migrant population would be covered by a survey which restricted its fieldwork to wards at least as concentrated as those shown in the first column. For example, if fieldwork in England was restricted to wards in which the authors estimate that at least 25 per cent of households contained migrants, one would expect the survey to cover 38 per cent of the English migrant population.

On the basis of the current priorities for a survey and potentially available resources, the authors recommend limiting population coverage to 80 per cent of both English and Scottish migrant populations, and in the remaining discussion the authors have assumed that this cut-off will be applied. This seems a reasonable compromise between the need to cover most of the migrant population and the inevitable cost limitations. This would enable fieldwork to be restricted to the 28 per cent of wards in which five per cent or more households contain migrants in England and to the 38 per cent of wards in which two per cent or more households contain migrants in Scotland (migrants are less concentrated in Scotland than in England).

The authors recommend that a cluster sampling approach be used by default in the interests of cost-effectiveness, and that wards are suitable for use as clusters because the small area estimation methods underlying the sample design (see section 4.1) are based upon these. Simply stated, this means that the sample design would involve (i) drawing a sample of wards and (ii) drawing a sample of addresses within sampled wards (however, in areas in which ACS or Waksberg methods are used further clustering – at the level of the 'block' – will take place within wards).

If migrant addresses are selected with equal or near equal,¹² probability, and if a set of assumptions listed below were to prove to be true, the authors estimate 82,500 addresses would need to be issued for screening in order to achieve the assumed target of 6,000 interviews in England and Wales and to issue about 125,000 addresses for screening in order to achieve 3,000 interviews in Scotland. The assumptions upon which the authors have based these estimates are (i) that their estimates of numbers of migrant households are accurate (ii) that the sample will be drawn from the Postcode Address File and that about 92 per cent of sampled addresses will be residential (iii) that 80 per cent of residential addresses will successfully be screened (iv) that one migrant interview per household will be conducted (v) that a 60 per cent response rate with identified migrants will be achieved and (vi) that face-to-face screening will be used throughout.

These levels of screening are high, and therefore costly. Three approaches may be taken to reduce the number of addresses requiring screening. The first would be to use more focused screening methods in areas of low concentration.

The second would be to allow more than one interview per household and/or to interview migrants living in more than one household at the same address. However, very little is generally known about the structure of migrant households, the approach would introduce household level design effects which might reduce the precision of survey estimates and would increase survey burden in households containing more than one eligible migrant and this may have a detrimental effect on response rates, and finally, there is a risk of 'contamination' whereby answers given by the first household respondent influence those given by another. On balance, the authors therefore advise against this approach.

The third approach would be to move away from an equal probability design. If a higher proportion of the achieved sample is taken from areas of higher migrant concentration, the average eligibility rate amongst screened addresses would increase and a smaller start sample would therefore be required. The disadvantages of this approach are that it would (i) diminish the number of migrants sampled from areas of lower concentration and (ii) lead to a requirement for additional weighting, which in turn would reduce the precision of survey estimates. It would, however, substantially reduce the number of addresses to be screened, and the authors suggest that these gains would be sufficient to render this approach attractive.

To give a realistic illustration, if in England a sample was taken of addresses in wards containing 20 per cent plus migrants at four times the rate at which addresses were sampled in less concentrated wards, the number of addresses issued for screening would reduce from 83,000 to about 57,500.

As a result of this survey, design effects would be increased by a factor of about 1.55. The authors have assumed that with an equal probability address sample the value of the survey design effects will be about two – as a result of clustering and respondent selection

12 The authors suggested hybrid ACS/ Waksberg method would not deliver exactly equal probability samples of addresses.

weighting. The revised approach just discussed would increase these to about 3.1. In turn, this means that the effective sample size would decrease in England and Wales from around 3,000 to about 1,935 and in Scotland from about 1,500 to about 970. In other words an equal probability address sample would deliver confidence intervals of similar width to those delivered by simple random samples of one half of the size of the actual achieved samples – about 3,000 in England and Wales and 1,500 in Scotland. And the unequal probability address samples would deliver confidence intervals equivalent in width to those delivered by simple random samples of 1,935 and 970 in England and Scotland respectively. Whilst an increase in design effects is clearly not ideal, it may be a serious consideration to help make the survey more cost-effective.

Sub-groups

As discussed above, ideally the survey should deliver robust estimates for a number of sub-groups. Unfortunately, given the lack of definitive information on the characteristics of migrants nationally or locally, for most sub-groups it is not feasible to estimate how many migrants of different types will be delivered by a sample of fixed size. Although the authors feel that their local area modelling should deliver reasonably robust indications of how many migrants live in particular areas, they do not permit disaggregation by migrant types.

Survey costs

The very rough estimates of indicative costs suggest that the costs of the two illustrative designs discussed above would be as shown below.

Table 4.5 Sample sizes and cost

Country	Sample size	Type of sample	Indicative cost
England	6,000	Equal probability	£1,800,000
England	6,000	Increase sample rate by factor of 4 in 20%+ wards	£1,500,000
Scotland	3,000	Equal probability	£2,100,000 ^a
Scotland	3,000	Increase sample rate by factor of 4 in 7.5%+ wards	£1,600,000

^a Migrants are less concentrated in Scotland than England, hence a higher cost for a smaller achieved sample – most of the cost comprises screening addresses rather than carrying out migrant interviews.

In the purely statistical sense (i.e. ignoring costs) the equal probability sample represents the ideal. However it is more expensive and for this reason the authors would recommend some oversampling of migrants in more highly concentrated areas. That said, the figures above indicate that even with the unequal probability design costs are likely to be high.

Further reductions to the cost of a survey could be made by:

1. reducing population coverage from the 80 per cent level;
2. increasing the ratio of the sampling fractions used in areas of higher and lower concentration;
3. reducing the sample size.

The authors would advise against the first and second of these. Reducing coverage would mean that the survey would exclude a very substantial proportion of the migrant population which, in turn, would lessen the usefulness of the estimates generated. Increasing the ratio of sampling fractions across concentration bands would substantially reduce the statistical efficiency of the sample and would, therefore, widen confidence intervals.

The authors' overall recommendation would be to use higher sampling fractions in more highly concentrated areas as suggested above, and then to reduce the sample size to fit the available budget. If the four to one sampling fraction ratio discussed above was maintained, the indicative costs for samples of 5,000 and 4,000 in England and for samples of 2,000 and 1,000 in Scotland would be as follows:

England

5,000	£1,300,000
4,000	£1,100,000

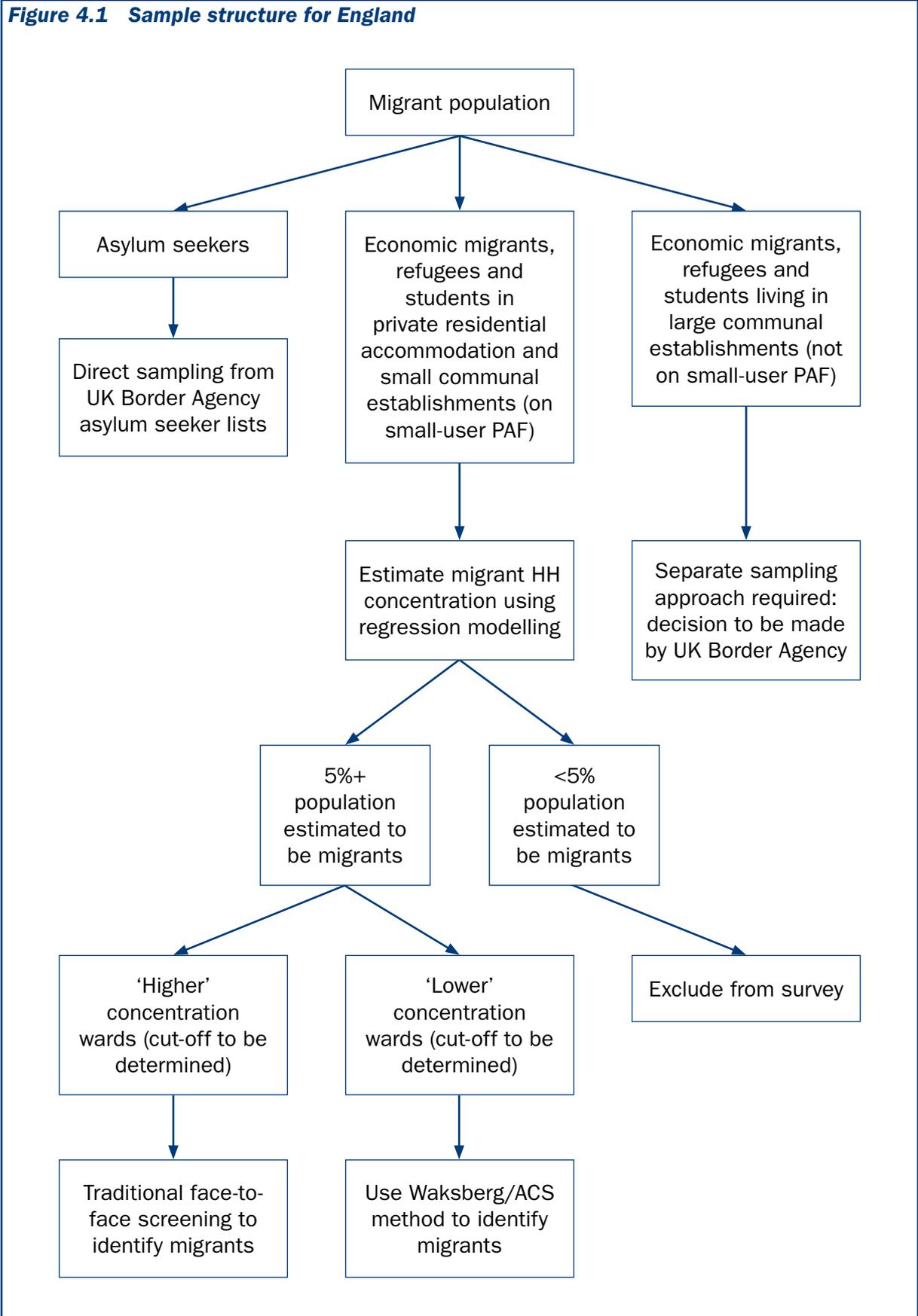
Scotland

2,000	£1,200,000
1,000	£750,000

Although as an overall approach this approach is recommended, it would be important only to finalise the design details after holding further discussions about exact survey priorities, discussions which should be informed by careful examination of how design choices will affect cost, effective sample size and likely bias.

Figure 4.1 summarises the proposed approach to sampling in England.

Figure 4.1 Sample structure for England



Appendix 1 Sampling asylum seekers

Sources for sampling asylum seekers

All asylum seekers in the UK are required to provide contact details to the UK Border Agency. This information is held in a 'caseload' database owned by the UK Border Agency. Hence, if this contact information is sufficiently accurate and up to date, it should be possible to sample asylum seekers directly.

Main caseload database

For this feasibility study, an extract of the asylum caseload data was generated by the UK Border Agency on 23 September 2009 containing name, address and nationality, and also information on type of claim, case start and end date and outcome type (where case finalised), and whether the applicant was claiming support. All dependants were removed, resulting in a file of over 12,000 cases. This included cases that were currently under consideration (i.e. asylum seekers) and also closed cases (those who had been granted refugee status – failed asylum seekers were not eligible to be included in the sample).

Support databases

A proportion of the addresses held related to supported accommodation provided by the Government or Local Council, while others were the addresses of relatives or friends of the claimants. It was felt that the supported accommodation claimants were more likely to be found at the addresses provided, but were also very unlikely to be at these addresses beyond a month after a decision on their claims. One requirement of the fieldwork tests was to check this assumption.

More detailed information on applicants in supported accommodation would need to be drawn from other databases such as those for supported accommodation claimants and for subsistence-only claimants. The supported accommodation database also includes applicant's date of birth, language spoken and support provider name, which could be useful for the survey (e.g. for tailoring advance letters) although the authors note that only a small proportion of this database could be matched with the main caseload database so this might not be useful and could lead to more pronounced non-response amongst those not included if special contact measures are devised for the survey based on language etc. The subsistence database contains basic contact information only. The supported accommodation database contains predominantly applicants aged over 18.

Unaccompanied asylum seeker children databases

A further potential source of information for the survey is the UASC databases held by Local Authorities. Data were provided for Leeds and Solihull after field tests in these areas indicated that a number of minors aged under 16 had been sampled from the caseload database. The utility of these data is explored further below.

The key requirement for combining these data successfully is a unique reference number common to them all. The Home Office reference number appears to provide this function, but has some limitations which hinders how well the supported accommodation and main caseload database can be matched.

Fieldwork tests

In total, just over 3,000 postcode sectors were represented in the data (there are around 8,200 in England and Wales), and 1,900 if outstanding (e.g. unconcluded) cases only are considered. Postcode sectors are of a similar size to wards, typically contain around 2,000 addresses, and are a fairly standard area size used for fieldwork as they are manageable areas for a single interviewer to cover. While it is possible to use larger areas, particularly with named and eligible sample which will mean that fewer addresses need to be visited to achieve an interview, this would have implications for costs in terms of travel time and expenses.

The average number of asylum contacts per postcode sector in the database was 5.1 overall, with 3.7 across decided/completed cases and 3.4 across undecided/incomplete. Looking at the undecided cases alone (i.e. asylum seekers rather than refugees/failed asylum seekers), just 112 postcode sectors contained ten or more contacts, and 374 contained five or more. However these 374 sectors (comprising 20% of all postcode sectors with pending cases) contained 57 per cent of the database contacts, indicating that it would be possible to focus fieldwork on the most concentrated areas at the expense of full coverage, and make the survey more efficient. This approach should therefore be considered in light of budget and requirements.

Fieldwork tests were conducted in two dispersal areas, Leeds and south Birmingham/Solihull, and the most concentrated postcode sectors within these geographies were selected to make the tests more efficient/expedient. Even so, around a dozen postcode sectors were required for each to produce sufficient potential asylum seeker contacts for the tests. Fifty addresses were issued for each area, with around 25 undecided cases and 25 decided cases in each area. For the decided cases, a range of recent and less recent decisions were sampled, up to a maximum of six months post-decision, in order to provide feedback on how quickly asylum seekers leave various types of accommodation.

Advance letters were posted to every address several weeks before making contact. These included the UK Border Agency logo and contact details. Most asylum seekers remembered receiving the letter and several called the office (or others did so on their behalf) in advance of fieldwork. As with the other methods, interviewers were briefed in person, and the pre-printed contact sheets contained additional feedback information concerning the type of accommodation found. Fieldwork took place in November 2009.

The importance of contacting social services in advance was emphasised by the numbers of asylum-seeking minors (aged under18) that were found in the field. These cases generated a raft of calls in response to the advance letters from concerned foster parents and social services. It would have been preferable had the databases been able to distinguish more

adequately which contacts were minors. This is something that could be explored before undertaking the main stage fieldwork of a survey.

Findings from the tests

The following table shows the overall results from the tests. Ninety-nine named contacts were issued to interviewers in total, and once vacant or derelict addresses were removed, no contact was possible with 55 per cent, 21 per cent were unproductive for other reasons and successful interviews were conducted with around a quarter (23%).

Table A1.1 Results from asylum field tests

	Total	Solihull	Leeds
Issued sample (named contacts)	99	49	50
Deadwood (derelict/vacant/not found)	5	2	3
Non-deadwood sample	94	47	47
No contact, of which:	52 (55%)	22 (47%)	30 (64%)
No contact at address	15	-	15
No contact with named respondent	7	4	3
Not known at address	13	8	5
Moved from address	17	10	7
Other unproductive, of which:	20 (21%)	13 (28%)	7 (15%)
Mother tongue interview required	7	3	4
Refusal by named contact	6	4	2
Refusal by gatekeeper/warden	7	6	1
Successful interview	22 (23%)	12 (26%)	10 (21%)

The tests had two main aims (i) to establish whether the contact details for asylum seekers are accurate and (ii) whether asylum seekers could be interviewed at these addresses. The tests affirm both of these aims.

- In total, only 18 out of 99 of the contact names/addresses were found to be inaccurate (e.g. deadwood addresses and those where the respondent was not known at the address, although the latter may not be correct as is dependent on the knowledge of others at the address). Of the remaining 81, accuracy could not be established for 15 (no contact at address) and the remainder (64 or 65%) were established to be correct. It is possible that a more intensive fieldwork effort would have enabled contact to be established at some of the 15 unknown addresses, and so one might expect the useable proportion to be higher than this.

- With respect to whether interviews might prove possible, a number of the unproductive respondents might have been converted in a more intensive fieldwork exercise. Efforts could have been made to pre-warn gatekeepers, other languages could have been catered for, and efforts could have been made to follow up movers (interviewers attempted to collect addresses for those who had moved, and succeeded with 8 out of 17 of these cases).

In summary, a survey of asylum seekers with sample of named contacts drawn from UK Border Agency databases is certainly feasible although there are likely to be problems attaining a high response rate. It would be beneficial if the various reference numbers contained in the databases could be clarified for matching purposes. It would also be beneficial if further, more comprehensive data which would enable identification of minors, could be obtained.

The findings also show that a survey requires extensive preparatory work to engage agencies and relevant community groups and enable access to asylum seekers, adequate language cover, and following procedures for those who have moved. This is particularly the case for social services, in the event that some UASCs may be contacted (whilst only those aged 16 and over would be eligible to participate, it is possible that younger applicants may inadvertently be contacted depending on how well minors can be identified).

Based on the tests it is estimated that around two-thirds of the sample could be valid/potentially productive, which suggests that a response rate in the region of 40 per cent would be possible. However, the tests were of a small size and in just two dispersal areas, and these numbers would not necessarily apply to a wider study. Cost/ efficiency/coverage trade-offs would also need to be considered in light of limited geographical clustering of the database addresses.

Appendix 2 Approaches to sampling that rely on social networks

Multiplicity sampling

Multiplicity sampling with rare populations relies on the fact that members are often interrelated by geographical closeness or other social factors. For example, if you sample me you can collect information on my sisters. Note that in the literature it is often expected that the single respondent tells you about the other members of the network, whereas in this case, the respondent would be used to identify other members who would then need to be interviewed. The theory owes much to the work of Sirken (for example Sirken, 1970). The network is generated by a rule linking respondents, such as siblings, close friends or neighbours. The link between respondents must be reciprocal and checked at interview. This reciprocal nature makes it possible to work out the multiple routes of selection.

This approach relies on migrants having strong local interconnections so that finding one migrant during some kind of screening process identifies the rest of their network for interview. One potential problem here is lack of knowledge regarding the size of people's networks. There are also issues around whether individuals feel comfortable providing contact details for other people they know. A related method – respondent driven sampling – might however assist with this issue. It requires that respondents recruit members of their network, rather than the interviewer. (See work by Heckathorn and others, e.g. Salganik and Heckathorn, 2004, for further details of this approach).

Field test of multiplicity sampling

Method and results

Multiplicity sampling field tests were conducted in two wards:

- East Ham, Newham (September 2009);
- Farley, Luton (September 2009).

These wards were purposively selected to have medium to high concentrations of migrants living in them.

Two hundred addresses were sampled in each ward from the post office small user address file. These addresses were screened by interviewers (face to face) for the presence of eligible migrants. At addresses where migrants were identified:

- short interviews were attempted;
- respondents were asked a series of questions designed to generate further sample members using the multiplicity sampling methods described above.

Addresses were issued in blocks of five adjacent addresses; each block was selected systematically from the list of addresses in the ward. Two interviewers worked on the test one in each ward. Up to two interviewer visits were made at each initial address before it was recorded as a non-contact, a low number for a full survey but all that was feasible for this test.

The table below shows the fieldwork outcomes for the initially sampled addresses. Screening interviews were completed at 61 per cent of non-deadwood addresses in Newham and at 72 per cent in Luton. Of the 50 eligible addresses in Newham, interviews were achieved at 25 (50%). In Luton eight interviews were achieved at 11 addresses identified as eligible (73%).

Table A2.1 Multiplicity sampling outcomes

	Newham	Luton
Issued addresses	200	200
Additional dwelling units contacted	15	
Deadwood addresses	4	8
Non-deadwood addresses	211	192
Number of addresses screened	128	139
Screening rate	61%	72%
Total no. of eligible households	50	11
Penetration	39.1%	7.9%
Achieved interviews	25	8
Interview response rate	50%	73%

Following each successful interview the interviewers attempted to establish how many eligible migrants living in the same ward were known by respondents or by other members of respondents' households. The interviewer attempted to recruit all eligible migrants so identified.

Two recruitment methods were attempted. In Newham interviewers attempted to recruit new sample members directly, whereas in Luton respondents were asked to recruit on the interviewer's behalf. The methods are described further below.

Beginning with 25 and 8 eligible interview respondents (see table above) in Newham and Luton respectively, interviewers asked the following question at the end of the interviews.

For this study we would like to talk to as many people as we can who, like you, are or have been at some time a citizen of another country.

Do you know any people in the local area who have ever been citizens of another country? By local I mean within about 1 mile or 1.5 km of here.

AND IF OTHER ADULTS LIVE IN THE HOUSEHOLD

Does anybody else in your household know any adults aged over 16 living in the local area who have ever been citizens of another country?

Seven (28% of 25) respondents in Newham and three (38% of 8) in Luton confirmed that they knew at least one such person. Interviewers then asked the respondents how many such people they knew.

Now I'd like you to think about all the adults aged over 16 who live in the local area who have ever been citizens of another country who you know or who other people in your household know. How many of these people do you or other members of your household know?

In total, 11 and 3 potential new sample members respectively were identified at this stage, 14 in total. The interviewers then asked a series of questions about each of these individuals, designed to establish their eligibility. The eligibility requirements are listed below.

The potential sample member needed to live in the same ward as the respondent – a map was shown for this purpose. A total of 6 of the 14 potential sample members in the two areas were excluded at this point because the contacts lived outside the ward boundaries.

The respondent was required to have visited the potential sample member's address or vice versa – the method requires the relationship between respondent and contact to be symmetrical and unequivocal, and it was felt that the chosen home-visit link would be less open to differences in interpretation than say one chosen on the basis of unqualified friendship, and yet more productive than one chosen on the basis of kinship (e.g. siblings, other relatives). No migrants were ruled out by this requirement.

Procedures for following up identified eligible migrants differed in Newham and Luton. Respondents in Newham were asked to provide the contact details (name, address) of all eligible migrant contacts. Interviewers were then instructed to send an opt-out letter to identified contacts and then to attempt interviews a week after posting this (assuming the sample member had not opted out).

Respondents in Luton were asked to send opt-in letters to eligible migrants themselves and, ideally, to follow this up and encourage contacts to take part (based on the principles of respondent driven sampling as mentioned earlier).

In both wards, respondents were paid an incentive of £5 in vouchers to identify eligible migrants, and were also promised an additional £5 for each interview achieved with eligible migrants so identified. Eligible migrants identified through these procedures were offered £10 in vouchers if they completed the interview. This approach was to test whether small incentives could encourage both parties to engage in the survey.

Of the six identified migrants living in the same ward as initial respondents in Newham, addresses were provided and opt-out letters were posted to four contacts. Interviewer follow-up established that only two of these four addresses were accurate and an interview was achieved at one of them.

Of the two identified migrants living in the same ward as initial respondents in Luton, an opt-in letter was posted to one contact; no reply was received.

The following table summarises what happened with each of the 14 potential new sample members.

Table A2.2 Outcomes for new sample members from multiplicity sampling

	Newham	Luton
Number of migrants for which household grid completed	11	3
Ineligible, live outside the ward (map check)	5	1
Migrants with eligibility confirmed	6	2
Address known, opt-out/in letters posted	4	1
Of which: address vacant/doesn't exist	2	N/A no response
Of which: away during fieldwork	1	N/A no response
Contact made with respondent	1	0
Interview achieved	1	0

Conclusions

The key criteria against which one should judge the success of the method are (i) whether it improves fieldwork efficiency by producing more migrant interviews, and (ii) whether it enables accurate information about the size of respondents' networks to be collected -- information that is required for the calculation of selection probabilities and design weights. In both respects the method failed – most notably for its lack of success in generating additional interviews; just one was achieved. Furthermore this additional respondent said he/she did not know the person who had identified him/her.

The method also appears suspect in other respects. While few respondents refused to go through the exercise outright, a majority said that they did not know any eligible migrants. The interviewers were not convinced that respondents were telling the truth about their networks but that they were unwilling to provide information about others they knew.

The authors therefore conclude that multiplicity sampling under these conditions would be unworkable for a future large-scale random probability survey of migrants, and that even non-probability methods that recruit via social connections in the population would likely struggle.

Appendix 3 Approaches to sampling that rely on localised clustering

Adaptive cluster sampling (interviewer identifies)

This is an area based approach that has been used to count wildlife populations. It has a sound theoretical basis as a random probability approach and as such can give unbiased estimates of a rare population. However, as with most sample designs, moving the theory seamlessly into the real world of non-contacts and non-response will require some thought. Think of a game reserve that has been split up into small geographic areas. You take a random sample of areas in the park and go and look for elephants. If there are none in that small area you stop. (Note that elephants cannot really hide during this process while humans in households and addresses can.) However, if there are elephants in the area you then move to all the adjacent areas and you keep going until you do not find any more elephants. This approach works well assuming elephants are relatively clustered in the park, so that having found a few in this area there are likely to be more in the surrounding areas; but unlike focused enumeration they do not need to know about each other. It results in a random sample of clusters of the population of interest and, therefore, allows design-based¹³ unbiased estimates of the population to be produced.

The selection of a sample of wards based on the frame information might mean that each ward could be thought to act like its own game reserve. Within wards one might then create blocks of five addresses (in some geographic order), then select a sample of blocks spread through the ward and screen for migrants. If one is found, then adjacent blocks are added and so on. Below is a pictorial example taken from Thompson (1990). The left half of the picture shows the initial selection of ten address blocks for screening. Eight of those blocks contain no migrants but in two blocks migrants would be identified at screening and, therefore, moving to adjacent blocks until there are no migrants in a block, the two clusters of blocks would be formed as shown on the right-hand side.

¹³ Design-based refers to the properties of the estimator under replication of the sample design.

Figure A3.1 Example of adaptive cluster sampling

Thompson: Adaptive Cluster Sampling

1051

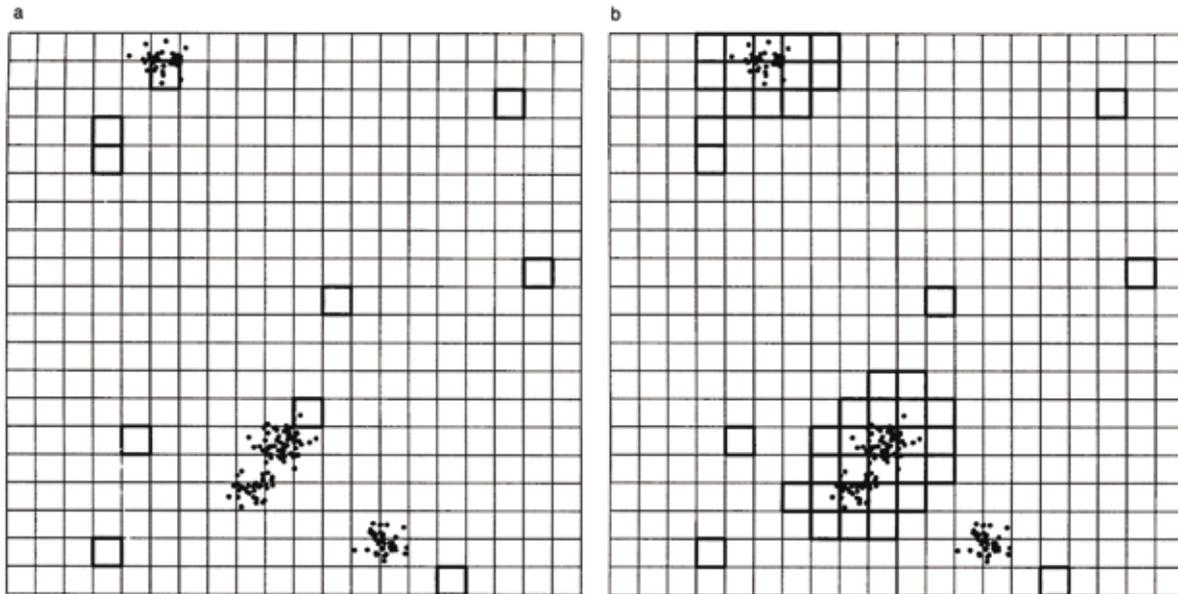


Figure 1. Adaptive Cluster Sampling to Estimate the Number of Point-Objects in a Study Region of 400 Units. An initial random sample of 10 units is shown in (a). Adjacent neighboring units are added to the sample whenever one or more of the objects of the population is observed in a selected unit. The resulting sample is shown in (b).

Field test of adaptive cluster sampling

Adaptive cluster sampling field tests were conducted in four wards:

- Bridges, Gateshead (July 2009)
- Lower Stoke, Coventry (November 2009)
- West End, Dundee (November 2009)
- Fountainbridge, Edinburgh (December 2009/January 2010)

These wards were purposively selected to have medium to high concentrations of migrants living in them. The initial fieldwork test was conducted in Gateshead using blocks of ten addresses and 200 addresses in total. Later tests used smaller blocks of five addresses and fewer addresses overall (50 per ward).

In all tests, blocks of addresses were sampled using the method of fixed interval and random start.

In Gateshead and Dundee a maximum of two calls at each address were possible in the time allocated for fieldwork. In the other areas it was possible to make up to four calls at each address in the allocated time (this level is closer to what might be done in a main survey, and hence gives a better indication of likely outcomes).

The table below summarises the fieldwork outcomes for all addresses screened-- both initially sampled addresses and addresses identified in the blocks subsequently are brought into the sample.

Table A3.1 Outcome from Adaptive Cluster Sampling

	Gateshead	Coventry	Dundee	Edinburgh
Interviewer calls per address	2 calls	4 calls	2 calls	4 calls
Issued addresses	200	50	50	50
New block addresses	210	70	0	45
Deadwood addresses	18	9	4	2
Non-deadwood addresses	392	111	46	93
Number of addresses screened	217	86	33	81
Screening not attempted (block eligibility already established)	38			
Screening rate	62%	77%	72%	87%
Total no. of eligible households	29	12	0	7
Penetration	13.4%	14.0%	0.0%	8.6%
Achieved interviews	17	11	0	7
Interview response rate	59%	92%	N/A	100%

In Dundee no eligible migrants were identified in the initially sampled blocks, and as a result no new blocks were brought into the sample. The area of Dundee selected contained both university halls of residence and a medical training facility with a campus. It therefore seems likely that the selected ward did indeed include many migrants, but that these individuals lived in properties which were inaccessible to the field tests (e.g. in halls of residence, medical institutions).

Interviewers were able to apply the method with relative ease -- it is essentially a standard face-to-face screening exercise, but with the added element of adding new addresses according to pre-specified rules. However, it became apparent early on that a major potential difficulty which would need to be overcome when applying the method in practice would be setting a block size which generated a sufficient, but not excessive, number of interviews. In order to follow pure random probability principles, it is essential that the method is followed to the end, i.e. until the interviewer reaches the point where every block or blocks containing eligibles is enclosed by blocks on either side containing no eligibles. The point at which this happens depends on two linked factors – the local concentration and distribution of migrants and the choice of block size. If the blocks are too small for a given migrant concentration, then one risks generating no/too few migrants. If the blocks are too large it is likely that they will continue to generate new blocks until the entire ward, or a very large part of it, has been screened; this would result in (i) unacceptably large clusters of interviews, with resultant damage to the precision of survey estimates and (ii) a protracted fieldwork period (in a typical ward of around 2,000 or more addresses, around 10 to 20 iterations would be required to screen the entire ward).

Of the test areas only Coventry, Dundee and Edinburgh, where blocks of five addresses, were used, were fully worked until no new migrants were identified (in Dundee this was true only in a trivial sense because no migrants were discovered in the initial sample of addresses). In Gateshead with blocks of ten addresses, screening was stopped once a second iteration was complete. In this area the initial 20 blocks opened 22 new blocks which opened a further 24, and it therefore appeared likely that the assignment would have continued to increase in size with each new iteration.

It is worth noting that in the multiplicity test areas addresses were also sampled in blocks of five addresses, and one can therefore estimate for these areas how many blocks would have been generated after a single wave of fieldwork, had adaptive cluster sampling methods been employed. Here the Newham ward appeared to be too concentrated even with a block size of five – the initial 40 blocks would have opened a further 54. However, in Luton 18 blocks were opened from the initial 40, suggesting that it may have been possible to work this area to a conclusion under the method.

The findings for this method are therefore somewhat mixed. As the figures in the table below show, the rate of identification or number of additional interviews achieved at new addresses is positive if not hugely impressive. The figures suggest that the method is indeed identifying clusters of migrants, but that such clustering is not very pronounced.

Table A3.2 Adaptive cluster sampling screening rates in original addresses and newly generated addresses

	Gateshead	Coventry	Dundee	Edinburgh	Total
Addresses issued & not deadwood	190	43	46	48	329
Interviews achieved (out of original addresses)	9	3	0	2	14 (4.3%)
New addresses generated which not deadwood & contact attempted	164	68	0	45	277
Interviews achieved (out of new addresses)	8	8	0	5	21 (7.6%)

Appendix 4 Glossary

A2	Bulgaria and Romania that joined the EU in 2007.
A8	the A10 countries excluding Cyprus and Malta (as these were already members of the Commonwealth).
A10	the ten countries (Poland, Hungary, Slovakia, Latvia, Lithuania, Czech Republic, Slovenia, Estonia, Cyprus, Malta) that joined the European Union (EU) in 2004.
DCSF	Department for Children, Schools, and Families (now DfE Department for Education).
DWP	Department for Work and Pensions.
HESA	Higher Education Statistics Authority.
IMPS	Improving Migration and Population Statistics (a project based within ONS).
LLMDB	the Life-time Labour Market Database is an administrative dataset linking the records of one per cent of National Insurance numbers over time.
ONS	Office for National Statistics.
PBS	the new 'Points-Based-System' that is being implemented for those coming from non-EU countries (referred to as 3rd Country Nationals).
PLASC	Pupil Level Annual School Census, an individual record of each pupil registered that year, which can be linked over time.
WRS	the registration system run by UK Border Agency for new employees from the A8/A2 countries.

References

- Brown, C. and Ritchie, J. (1981) *Focussed Enumeration. The Development of a Method for Sampling Ethnic Minority Groups*. London: Policy Studies Institute and Social and Community Planning Research.
- Kalton, G. (2009) 'Methods for oversampling rare sub-populations in social surveys'. *Survey Methodology*, 35, pp125-141.
- Kalton, G. and Anderson, D. W. (1986) 'Sampling Rare Populations'. *Journal of the Royal Statistical Society A*, 149, pp 65-82.
- Office for National Statistics (2006) *Report of the Inter-departmental Task Force on Migration Statistics*. <http://www.statistics.gov.uk/about/data/methodology/specific/population/future/imps/updates/downloads/TaskForceReport151206.pdf>
- Pickering, K., Smith, P., Bryson, C. and Farmer, C.(2008). *British Crime Survey: options for extending the coverage to children and people living in communal establishments*. <http://www.homeoffice.gov.uk/rds/pdfs08/horr06b.pdf>
- Salganik, M.J. and Heckathorn D.D (2004) 'Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling.' *Sociological Methodology*, Vol. 34, pp. 193-239.
- Sirken, M. G. (1970) 'Household Surveys with Multiplicity'. *Journal of the American Statistical Association*, 65, pp 257-266.
- Smith, T. M. F. (1983) 'On the Validity of Inferences from Non-random Sample'. *Journal of the Royal Statistical Society A*, 146, pp 394-403.
- Thompson, S. K. (1990) 'Adaptive Cluster Sampling.' *Journal of the American Statistical Association*, 85, pp 1,050-1,059.
- Waksberg, J. (1978) 'Sampling methods for random digit dialling'. *Journal of the American Statistical Association*, 73, pp 40-46.

