METHODOLOGICAL ISSUES RELATING TO HOUSEHOLD AND MIGRATION SURVEYS

Unpublished paper delivered at the HSRC migration workshop Pretoria 17-20 March 2003

J.A. van Zyl
HSRC
March 2003

INTRODUCTION

This paper provides a short overview of methodological aspects of household surveys in general, and migration surveys in particular. Where possible, the discussion emphasizes such aspects within the South African context. The main aim of the paper is to provide a framework to evaluate the methodologies used in the HSRC's "Causes of Migration" survey.

DISTINCTION BETWEEN HOUSEHOLD AND MIGRATION SURVEYS

Household sample surveys are conducted to collect information from households and/or individuals in a population. Depending on the topic under investigation, such surveys may want to collect factual data, information on the attitudes of people, their behaviour patterns, levels of knowledge, etc. Migration surveys can be seen as a sub-set of general household surveys. In general, similar methodologies will be used when conducting a migration survey, although slight adaptations to the methodology may be required. Household surveys have to deal with many problems, either in terms of design or actual conduct, which also apply to migration surveys. Obviously migration surveys will be facing specific problems as a result of the topic under discussion.

Within the broad field of migration, distinctions are made between international and internal migrants, (crossing of national borders versus the crossing of internal borders and boundaries), between short-term and permanent migrants, (using a temporal dimension), between documented and undocumented migrants (introducing a legal dimension), between voluntary migrants and refugees (introducing the element of choice), between long distance and short distance migrants, (introducing a spatial dimension), between migration and residential mobility, (introducing a definitional aspect) etc. These are only a few of the distinctions that can be drawn in the field of human migration. Using alternative migration typologies, it is possible to define other categories. The topic under investigation and the research questions will also influence the preferred methodology to be used in a study.

It is useful to consider the reasons for collecting migration data or conducting migration studies before attempting to describe or list the methodologies used in sample surveys.

Research questions and data needs

The purpose of a specific research endeavour, and its underlying questions, will determine the type of data needed for its execution. Thus, by implication, the purpose of the study, will also determine the methodology employed to collect such data.

i) If the aim is to obtain mere numbers of international migrants (with a limited number of variables e.g. age, sex and country of origin), use will be made of administrative data collected at border posts of a country and published in a specific format. However, if the data is not complete, consideration has to be given to other sources and methods.

ii) If the aim is to obtain information on migrant proportions (e.g. the proportion of the foreign born in a population or even duration of stay), use can be made of population census data.

iii) If the study aims were the motivational aspects or the reasons for migration, a typical approach will be to design a questionnaire and use a sample survey to elicit relevant information. Population census data would not be very useful in this regard because census questionnaires are limited in their ability to accommodate detailed questions on specific topics. Another approach would be to utilize a longitudinal or panel design.

iv) If the aim of the study were to analyse the selectivity of migration, it would be possible to use either census data or survey data (provided that migration questions have been asked).

Type of household sample surveys

Cross-sectional surveys

Household sample surveys are conducted with the aim of collecting information at a specific point in time. The design of the survey provides a cross-sectional view of a specific population. The data from such surveys is used to analyse a variety of topics and the findings can be compared with similar information collected in different areas or at

different times. By drawing successive samples from the same population, the data form a time series and should not be confused with longitudinal data. The data collected by the main survey of "The causes of migration" study is typical of data in cross-sectional surveys.

Panel surveys

A panel survey is designed with a view to re-visit the same sampling points (households) covered in a first/baseline survey. The results of subsequent surveys provide the analyst with more dynamic information than would be possible with a cross-sectional survey. This design also provides the opportunity to collect more relevant migration data.

Longitudinal surveys

At the heart of the design of a longitudinal study lies the intention to follow households/individuals over time, even if they move away. This is the essential difference between a longitudinal and panel study. Such follow-up activities are complex in nature and attrition of the original sample of the longitudinal survey is a methodological issue. A longitudinal approach in migration research can become a powerful instrument to compare migration intentions and actual migration behaviour.

The design of "The causes of migration" incorporated a number of features to conduct either a longitudinal or a panel survey. For instance, the detailed maps and listing information compiled by the fieldwork companies enabled the HSRC researchers to return to the same households after the original visits. This information is crucial if a follow-up study is planned in a couple of years. The questionnaire also included the address details of family and friends of respondents to assist in locating a household/individual in the event of a move. During the main survey, a number of respondents expressed reservations about providing this information that many regarded as confidential.

ERRORS IN HOUSEHOLD SAMPLE SURVEYS

The reason for considering the kinds of errors that occur in household sample surveys is two-fold. On the one hand it indicates possible areas in collected data that maybe error prone. More importantly, the awareness of such errors serves as a catalyst to improve methodologies.

Two major types of errors occur in surveys. The first can be described as sampling errors and occur because the results of the findings are confined to a sample of the population, rather than the whole population. The second type of error can be described as

non-sampling errors. Non-sampling errors can be classified into three major categories, namely coverage, non-response and response errors (United Nations1982).

Coverage errors

Coverage errors refer to the failure to include all potential respondents in the sampling frame.

Non-response errors

Non-response errors occur when households or individuals selected for inclusion do not provide all or only partial information. Total non-response is the failure to collect any information, while partial or item non-response refers to the failure to collect specific items of information (but only taking into account those who should have responded to a particular item).

The total non-response rate is calculated by comparing the units selected for the sample with the units returned from the field. In the "Causes of migration" survey non-response differed markedly by population group. Whites were particularly hard to contact or difficult to interview.

Table 1: Non-response by population group*

Population group	Non-response rate* (%)
African	9
Coloured	18
Indian	11
White	30

*Unweighted

As an example of item non-response, reference is made to data collected during the "Causes of Migration" survey. Table 2 provides detail of response to the question of monthly income by type of migrant. Of interest is the large percentage of international migrants, especially cross border migrants, who did not provide any income detail.

Table 2: Item non-response: Monthly income

Migrant category	Non-response rate* (%)
Non-migrants	3
Internal migrants	6
Cross border migrants	43
Other international migrants	28

*Weighted

When considering response rates the following aspects deserve attention: Whether or not to use weighted non-response rates, household versus individual non-response, whether responding units differ from non-responding ones (the characteristics of non-respondents), the sources of non-response and how can it be reduced (United Nations 1982:58). Regarding the latter, response burdens and the role of the interviewer can be highlighted.

Total non-response is mainly due to non-contacts and refusals. Non-contact includes failure to locate a specific sampling unit, or respondents not being at home. Refusals may occur due to the topic of a survey. In migration surveys respondents may view the survey as an attempt to enquire about their legality in an area. Such an attitude could impact on refusal rates. The role of interviewers in refusals may also be important in certain circumstances.

Response errors

Response errors refer to data that has been obtained, but is incorrect. Sources of error include respondents not knowing the answer, problems of recall and respondents purposefully providing wrong information.

METHODOLOGY OF HOUSEHOLD SURVEYS

Household sample surveys, as a tool to collect information, have been used in many disciplines (United Nations 1971; Warwick and Lininger 1975, Babbie 1973, Shryock and Siegel 1976, Marsh 1982, Bilsborrow, Oberai & Standing 1984, de Vaus, 1986, Bogue, Arriaga and Anderton 1993). Over time the methods and procedures used to carry out household sample surveys have been developed and refined. As an area of specialisation, sampling underwent a similar development (Hansen, Hurwitz and Madow 1953, Kish 1965, Levy and Lemeshow 1999). The overall aim of these improvements was to assist in the collection of more accurate data. Areas of improvement included sampling design, selection of sampling units, instrument/questionnaire design and fieldwork methodologies. In fact, during the past three decades, major developments occurred on how sample surveys are conducted, as well as in the analysis of the results of sample surveys. Examples that come to mind are telephone sampling and interviewing methods, including computer assisted telephone interviews, and the use of computer-assisted personal interviewing techniques. Recently the HSRC conducted an HIV prevalence survey in South Africa (Shisana & Simbayi 2002). The survey applied a relatively novel approach to select an unbiased sample of respondents. Interviewers visited the selected households and completed a short household schedule, noting important demographic characteristics

of the members. These questionnaires were returned to the HSRC and the data captured. A random selection programme was written to select respondents according to three broad age groups. Another group of interviewers then visited the same households to interview the selected respondents. The main disadvantage of this approach is costs and the accompanying logistical complexities. Regarding the analysis of data, more user friendly computer software has become available to estimate standard errors and to carry out statistical procedures such as contingency table analysis and multiple or logistic regression that take into account complex sample designs. Advances in mapping and geographic related software have the potential to enhance sampling procedures in the field.

Adherence to known methodologies is advisable at all times. However, the mere adherence to such methodologies will not ensure perfect data, but known and tested procedures will go a long way to ensure a good survey. When conducting a scientific sample survey, it is assumed that such a survey will strive to incorporate all important methodological considerations.

However, if an inappropriate methodology is chosen or research methodologies are applied incorrectly, this will in all probability result in incorrect data and lead to erroneous findings. In the field of migration research, the HSRC undertook a study that went awfully wrong. The purpose of that study was to estimate the number of "illegal immigrants" in South Africa. Unfortunately, the methodological approach was unsuited for this purpose. The findings were quoted and used extensively without any critical appraisal (see Appendix 1).

The methodological elements of household sample surveys can be categorized as follows (also see for instance United Nations 1971, Warwick and Lininger 1975) Best practice require that when planning, preparing and conducting a survey, it will include most, if not all, of the following steps and activities.

Pre-survey activities/preparation

Formulation of research questions/proposals

Review of the literature/theories

Design of a draft questionnaire

Sample design

Listing/demarcation

Pilot survey

Finalisation of questionnaire/translations

Logistical arrangements

Conduct of the survey

Sample selection

Interviewer selection and training

Interviews

Checking of questionnaires

Back-checking/quality control

Data processing and analysis

Coding/numbering

Data entry

Computer editing

Weighting

Tabulation and analysis

Dissemination activities

Report writing

METHODOLOGY EMPLOYED BY THE CAUSES OF MIGRATION SURVEY

Given the fact that a considerable amount of time will be devoted to analyse and discuss the results of the "Causes of migration" dataset at this workshop, it is useful to consider the methodologies used in the study. This overview will not attempt to describe the methodologies in full (see Kok, Wentzel & Pietersen 2003) but rather to evaluate and critique selected aspects of a methodological nature. This can serve as an additional source to evaluate the usefulness, relevance and accuracy of data collected by this major migration survey.

Planning and preparation

The main survey was only conducted after a period spanning more than three years, during which planning and preparation was done. This included a thorough proposal to the HSRC, consideration of theoretical issues and a preliminary analysis of secondary data (see Kok et al 2002). From its inception, this study had a definite focus, i.e. looking at the causes of migration.

Initial survey

In the year 2000, a relatively large initial survey was conducted as part of the broader study. The main purpose of this initial survey was to evaluate the validity and potential reliability of scale items. The sample size of the initial survey consisted of 1 000 households. This survey incorporated basic procedures, e.g. a scientific drawn sample, a process during which a questionnaire was designed and translated and a small pilot survey to test the questionnaire. The survey itself was conducted according to accepted fieldwork practices. Upon completion of the initial survey, a provisional analysis of the data was carried out and reported in an unpublished manuscript (Kok and O'Donovan 2001). The initial survey results were used to identify and select a minimum core set of appropriate scale items.

Main survey

The main survey was preceded by a workshop attended by members of the project team. During this workshop the results of the initial survey was considered, as well as other evidence assembled during the planning phases of the project (for example an analysis of the 1996 census data regarding migration patterns). This workshop thus functioned as the final planning activity of the main survey. The next step was to finalise the questionnaire to be used in the main survey. The design was largely based on the questionnaire used in the initial survey, sans those items deemed not significant for the construction of scale items. A small pilot survey was nevertheless carried out to ensure that this redrafted instrument met minimum expectations in terms of length and understandability. Adaptations were made to the questionnaire based on the findings of the pilot survey. The instrument was subsequently translated from English into the main other languages in the country. Back-translation was used to ensure the correctness of these translations. A complex stratified sample was drawn using enumerator area data of the 1996 census. A cluster size of only 6 households per enumerator area ensured that sample design effects would be kept at a minimum. The size of the sample of more than 4 000 households aimed to ensure an adequate number of cases for analytical purposes. Selection of households and respondents were done in the field according to a set of guidelines. Fieldwork companies were obliged to obtain maps of each enumerator area and a complete listing had to be done in the field in each of the selected enumerator areas. Based on the number of dwellings in each enumerator area, a systematic sample was drawn to select the six ultimate sampling units.

Since the HSRC does not have its own fieldwork unit, fieldwork activities was contracted to two main fieldwork partnerships. Training of the fieldworkers was scheduled as formal training sessions. Researchers of the HSRC attended the majority of these training sessions in different locations in the country. The role of the researchers was to observe the training, but also to be available to answer any questions.

Fieldwork commenced according to schedule by following the agreed upon procedures. Although the fieldwork companies undertook to do quality control by revisiting/back checking, the HSRC decided to conduct an independent re-interview survey1. The contracted fieldwork companies were provided with details of a sub-sample of the original sample. The companies had to provide the HSRC with the completed questionnaires for these EAs, in addition to maps, the record of the re-listings, and details of the sampling procedure followed in the field. Upon receipt of this material, HSRC researchers conducted a series of re-interview visits. The results of these visits showed large deviations from agreed upon procedures in many enumerator areas (see Appendix 2 for a summary of the main errors encountered during the series of re-visits). The work of the fieldwork teams in many areas were found to be unacceptable, which resulted in the re-doing of a significant proportion of the sample, either by the same companies in some cases, or another company that was newly appointed for that purpose. In passing it should be noted, that the part of the sample re-done by this new company, was completed using a computer assisted personal interview methodology (CAPI).

Upon completion of the fieldwork, questionnaires were re-checked at office, coded and numbered. Questionnaires were captured by means of batch-entry methodology (with the exception of those questionnaires completed by means of CAPI). The next step was the weighting of the data, taking into account the design of the sample and non-responses. Finally, a data set was released for analysis.

Questionnaire issues

Section 6 of the questionnaire is titled: **Last Migration (If Ever Migrated)**

Section 6.1 is titled: **Origin: Duration Of Stay**

1 Where HSRC project members were involved in other surveys, they tended to accompany fieldwork teams at the beginning of the survey to observe how the work was being done and be available for any queries. In this study a decision was made to allow the companies to finish the work before doing an evaluation exercise in the field. In their proposals, the fieldwork companies emphasized the quality of their work and did not indicate the need for any assistance from the HSRC.

Question 6.1.1 is titled: **Have you ever lived outside "this area" for a period of at least six months?** The respondent could answer "Yes" or "No" to this question

This question tried to summarise the migration experience of a respondent in a single question. This question was also asked during the re-visits. In a number of cases the answers provided in the actual survey (answer "no"), did not tally with answers given during the follow-up (answer "yes"). This discrepancy was discovered when a variation on a migration history question was posed. "Where were you born"? Since the answer to 6.1 was "no", it was a surprise to find out the respondents were in fact migrants. When questioned about this discrepancy, many respondents said that they had understood the question to be: "Have you ever lived outside this area in the past six months".

One possibility is that the intended meaning of the question was lost in some versions of the translated questionnaires. In addition, by following a summarised approach to elicit a migration history, there is a risk of missing information because crosschecks cannot be made. Based on this observation, it is possible that the proportion of "migrants" may have been under-estimated during the survey.

What can we learn from this survey in terms of methodology

Mindful of current best practices in survey research, "The causes of migration in South Africa" survey was carried out. Nevertheless, an independent re-interview showed major problems in the data that was collected. Large parts of the survey subsequently had to be redone. This was a big disappointment for the researchers involved in the project,

However, had this independent re-interview survey not been done, we would have been unaware of flaws in the survey and would have been working today with much less accurate data. Quality control exercises of the kind that was conducted in the "Causes of migration" survey is very expensive, and the UN noted that because of this reason, this approach is not always followed. Although the quality control exercise was extensive, a part of the survey that was re-done, was not subjected to a second round of quality control visits.

Are there any lessons to be learned from what happened?

- Despite a whole set of guidelines and procedures being in place, it did not prevent the fieldwork activities from going wrong. However, had the study not followed a set of guidelines, the eventual outcome could have been much worse.

- From this experience it would appear that researchers who divorce themselves completely from data collection activities run a real risk of accepting sub-standard data upon completion of fieldwork. At best, the researcher(s) will have little or no idea of things that went wrong.
- Even trusted and respected fieldwork companies can be hoodwinked by clever fieldworkers. The example comes to mind of the interviewer, who knew that back-checks were being done by telephone, ensured that the one "real" interview had contact details. The other interviews in the cluster were "false" with no contact details. He knew that if control staff were able to confirm one interview in a cluster, there was a very good chance that all his other work will be accepted.
- Structural factors inherent in fieldwork procedures can contribute to data problems. For example, payment schedules that do not reward interviewers for not-at-home respondents, or refusals, may have the effect that interviewers substitute such households with other (non-selected) households/individuals. And if fieldwork companies place tight deadlines on the completion of the work, interviewers are in no position to wait and return later to look for hard-to-find households. They rather then make use of other short cuts.
- Sufficient training of interviewers is an essential element of good fieldwork. In those cases where the training was not as intensive (compared to other training sessions in the survey), it was noticeable in the significantly poorer performance of the interviewers.
- In the absence of good quality maps an/or photography, re-listing and mapping is essential to ensure that all households in an enumerator area have an equal chance to be selected. This procedure is invaluable when revisiting the area and to find the selected households. Subsequent to this survey, the HSRC developed a "master sample". Of particular interest is the use of aerial photography and Geographic Positioning Systems to identify individual dwellings. These tools will assist interviewers to locate the correct dwelling, even with the lapse of a number of years.
- If a quality control exercise is not done, a survey can never claim that the quality of the collected data is beyond reproach.

APPENDIX 1

THE HSRC FIGURES REGARDING "ILLEGAL IMMIGRANTS":

INTRODUCTION

This section is provided as an example of how a flawed methodology can impact on the findings of a migration study.

In the mid-1990s the Human Sciences Research Council (HSRC) released figures on the number of "illegal immigrants" in South Africa. Ranging from approximately a minimum of 2 million to a maximum of 8 million "illegal immigrants", these figures received wide media coverage and were often quoted. The reason for this was that no other estimates were available. At the time the HSRC estimates met with approval in some quarters and were severely criticised in others. The general gist of the criticism was that these estimates were wrong. In the face of persistent criticism, other researchers in the HSRC re-visited these calculations.2

ORIGIN OF THE HSRC FIGURES

Concomitant with the political changes in South Africa since the late 1980s, the country increasingly acted as a magnet for persons from other countries in search of employment opportunities and the like. Many of these persons merely crossed porous borders, or overstayed their visa expiry dates.

By the early 1990s, researchers who monitored public opinion in the former Centre for Socio-political Analysis of the HSRC became aware of increasing unease among the South African public regarding the number of foreign nationals in the country. This was established through regular public opinion surveys. The most often cited reason for this unease or xenophobia was the fear of job usurpation, whether such a fear had any basis or not.

These researchers subsequently set out to estimate the number of undocumented persons in South Africa. But persons who cross the national border of a country without the necessary documentation or overstay their visa would usually attempt to conceal their status, if not their presence. Such persons will therefore attempt to evade any form of perceived officialdom, lest they be identified as undocumented or illegal immigrants. Therefore, a population census would be unlikely to enumerate people who do not want to be counted. By the same token, all surveys would

2 These estimates of the number of "illegal immigrants" were formally retracted by the HSRC in 2002.

have difficulty in identifying such persons. In summary thus, one is faced by a situation of "*counting the uncountable*".

An analogy to this situation existed in South Africa before 1986 when African residents without legal residence rights in urban areas tried to avoid contact with officialdom, which was one reason for the systematic undercount in past population censuses.

Knowing this, these researchers decided to overcome the problem by making use of an indirect method. The questions were intended to shed more light on the vexed question of the size of the non-South African population. Since they realised that undocumented foreigners themselves would not provide accurate information, respondents in the survey were asked whether they knew of non-South Africans living in houses around their own, and whether they could provide an estimate of how many such people there were. The researchers drew up a set of questions to be included in the regular general-purpose surveys of the HSRC. Below is a verbatim description of the questions in the questionnaire as well as the findings of the December 1994 survey to these questions. Subsequent surveys in 1995 provided roughly similar findings.

1. During the past few months there have been many reports on illegal aliens in South Africa (e.g. Mozambicans, Nigerians, and Taiwanese). In your opinion should authorities:

Answer item	**%***
Act much more strictly against them	54
Act more strictly against them	15
Act less strictly against them	9
Act much less strictly against them	14
Uncertain/Do not know	9

*Weighted %.

2. Do any people who are not South African citizens live in the houses around this property?

Answer item	**%***	**N**
Yes	15,1	297
No	75,6	1637
Uncertain/Do not know	8,7	193
Not applicable	0,5	11

* Weighted %.

3. If yes, how many? .....

Number	%*	N
1	11,2	3
2	8,5	23
3	5,6	16
4	7,0	20
5	5,1	14
6	3,0	9
7	1,3	5
8	0,7	2
9	0,2	1
10	6,0	12
11	0,4	1
12	0,2	1
15	1,9	7
18	0,2	1
20	1,8	7
30	0,8	4
40	1,1	2
50	2,0	7
70	0,8	1
80	0,7	2
90	0,4	1
98	7,2	21
99	33,8	97
Total	100,0	284

* Weighted %.

Estimating the original number of "illegal immigrants"[3]

Approximately 15 % of respondents indicated that they knew of non-South Africans living in adjoining houses (see table on Question 2 responses). Of the 297 respondents, 284 gave a numerical answer (see Question 3 table). Of those, one-third was allocated a code "99", i.e. they did not know how many non-South Africans where living around them. Using the above data, the researchers then derived an estimate of the number of "illegal immigrants".

a) <u>The minimum estimate</u>

As far as can be ascertained the following argument was used to derive a minimum estimate:

It was assumed the respondents in the 15 % "knowing" households knew of at least one non-South African in the area adjacent to their dwelling. Fifteen percent of the adult population are resident in 15 % of the households. To extrapolate this to the total population, the researchers took the adult population in the country and multiplied that figure by 0,15 (that is, 15 %). They then multiplied this product by 1 (representing one non-South African). The answer was between approximately 2,1 million and 2,5 million.

[3]Our attempt to replicate the methodology in order to recalculate the exact original figures was not completely successful. We were unable to ascertain exactly the size of the adult population used in inflating the figures. An attempt to arrive at the original mean number of non-South Africans in the sample was also stifled, as we are unsure how the "don't know" responses were treated at the time.

b) The maximum estimate

To estimate the maximum number of non-South Africans, a mean was calculated of the reported number of non-South Africans (see table on Question 3). The product of the total adult population and 0,15 (the proportion who knew of non-South Africans) was multiplied by the mean number of non-South Africans (3,8 - 4,6). In this manner a total of more than 8 million non-South Africans were obtained.

Critique of the methodology followed to calculate the number of "illegal immigrants"

Although it was not possible to recreate the exact figures originally published, an examination of the methodology followed in deriving these estimates reveals fundamental flaws. They are discussed under three headings.

The data

To estimate the number of non-South Africans, respondents were asked "Do any people who are not South African citizens live in the houses around this property?" However, the number of "houses around this property" was seemingly not defined and the interpretation of this question probably depended on the trainer of the interviewers, the interviewers and the respondents themselves. The answer provided by respondents in some cases may have related to the nearest one or two houses, while in other cases, to all the houses in a block, or even to a larger area.

Another concern is that more than one household in the same cluster may have referred to the same non-South Africans living in the neighbourhood, leading to multiple counting of the same non-South Africans. In the sampling methodology, clusters of houses were visited to contain costs. A cluster usually consisted of a number of street blocks containing up to 150 houses in urban areas. In rural areas a cluster consisted of a village (or part thereof), or a number of adjacent farms. The ability of the respondents to identify non-South Africans was another concern.

The only conclusion one can reach is that the data assembled was of an unknown quantity, and certainly not suitable for extrapolation.

Method of extrapolation

The information collected during the surveys revealed that respondents in about 15 % of households knew of non-South Africans in "the houses around this property". In the original calculation the researchers utilised this fact, and proceeded to calculate that 15 % of all adults in the country knew of at least one non-South African (the minimum figure) and knew of approximately four (4) non-South Africans (the maximum figure). Two relatively simple examples will suffice to show the mistake that was made when using the inflation factor.

Let us assume there is a community consisting of 1 000 households. In total 2 500 adults reside in this community (2,5 adults on average per household). A survey is conducted among all the households and in each household one adult member is randomly selected as the respondent.

Example A

In the survey respondents are asked if they have a cell phone. Seventeen per cent of respondents answer that they have a cell phone. The question is how many cell phones adults in this community own. The answer is relatively straightforward. Multiply 0,17 (the proportion of respondents who have a cell phone) by 2 500 (the number of adults). The answer is 425 cell phones. We could do this because our respondents were a representative sample of adults.

Example B

In the same survey the respondents are asked if the household has a freezer. Seventeen percent of respondents answer they have one freezer. The question is how many freezers are in this community? The solution to this question is to multiply 0,17 (the proportion of households that have a freezer) by 1 000 (the number of households in the community). There are therefore 170 freezers in the community. In this case the freezers was used by households and not by individuals.

In our analogy, the method in example A was used in the original calculation of the number of non-South Africans, while it was more appropriate to use example B, as the household was the reference point. By using the method in example A, the number of non-South Africans were duplicated for other adult members of households, leading to a highly inflated number. And this does not even take into account the possible double counting of non-South Africans by respondents in the same sampling cluster.

Thus the conclusion can only be that the method to gross up the number of reported non-South Africans in the sample to a figure applicable to the country as a whole was wrong and led directly to the highly inflated figures.

The terms "illegal aliens" and "non-South Africans".

Respondents were confronted in the first question with the term "illegal aliens". The next question asked whether "non-South Africans" were living in the houses around this property. However, at that stage the questionnaire had already implied that these two terms are interchangeable, which they are not. A person can in fact be a non-South African and yet be legally resident in the country.

The questionnaire did not assist respondents to distinguish between foreigners with legal residence permits and those who had none, and this in fact contributed to more confusion. In reporting the findings, a quantum jump was made in the use of terminology. The data derived from the survey referred to the number of "non-South Africans" - irrespective of their legal status. However, this figure was reported as "illegal aliens". As it was not possible to distinguish between

"legal and non-legal" non-South Africans, it was inadmissible to use the term "illegal aliens" even if the number as calculated was correct, which in any event it was not.

Conclusion

The re-calculation of the "illegal immigrant" figures showed that the methodology used to calculate the original estimates was flawed. The estimates of the number of "illegal immigrants" were therefore also incorrect. The figures purporting to be the number of "illegal aliens" produced by the HSRC were inflated and cannot be used as an approximation, either for non-South Africans in a broader sense or for undocumented or otherwise illegal immigrants in a narrower sense.

What lessons can be learned from this exercise?

Data has limitations. Data should not be used for purposes it is not suited for. If, for example, it was reported that 15 % of respondents stated that non-South Africans where living in houses adjoining them, this would have been correct and beyond reproach. But by extrapolating the available data to estimate the number of "illegal immigrants" was unwarranted.

Researchers should not work in a vacuum but present and discuss findings and interpretations with colleagues in their own and other organisations. These figures were the result of one group of researchers working on their own. Had there been interaction with peers, it would have prevented these flawed findings being circulated. This saga also highlighted the need to make data and methodologies available to a wider audience. One element of scientific research is replicability. Everything possible should be done that this scientific requirement can be met under all circumstances.

APPENDIX 2

SPECIFIC FINDINGS OF THE HSRC QUALITY CONTROL SURVEY

This section contains a number of examples of the findings that were made during the independent re-interviews. HSRC researchers, who formed part of the migration project team, conducted the re-interviews. This is not intended to serve as a mere list of "fieldwork horrors", but rather to understand what went wrong and to assist in preventing the same mistakes being made again.

The "invention" of respondents

Probably the most serious example of fieldwork error occurred when interviewers "invented" households or respondents. A related phenomenon was when the re-interviewers could not find the households that were originally interviewed. Not all such cases were actual "fictitious" interviews. When a household could not be located during the quality control, it was classified as a "false" interview. What actually happened in the majority of cases was slightly more benign. It would appear that in certain cases, the interviewers, when finding no-one at the selected household, merely proceeded to interview another, unselected household, without altering the visiting details. For a re-interviewer, it is impossible to locate such a household without an address.

Why did interviewers resort to this tactic, which ultimately had dire consequences? The first was that interviewers did not want to return to the same household. In many cases this was a result of strict deadlines set by the fieldwork companies. Return visits, and time spent waiting for respondents is wasted time and cost money. In those circumstances interviewers devised their own strategies to overcome this problem. Another reason was financial rewards for the interviewers themselves. The interviewers were only paid for completed questionnaires. "Not at home" or "refusal" questionnaires did not offer the same rewards. For them, substitution was a viable option.

However, during the re-interviews, researchers did come across fictitious questionnaires. In the one instance, an interviewer was allocated a specific cluster. He would then proceed to interview the first selected household. If this household had a telephone, this was clearly indicated on the questionnaire. However, none of the other interviews in the cluster could be verified during the revisit.

What happened in this case? This was one of the most cynical forms of falsification I had personally heard of/come across during more than twenty years of fieldwork experience. The companies doing the fieldwork undertook to contact at least 10% of the original interviews to confirm whether an interview took place. For financial reasons, most companies conducted such checks by telephone. This particular interviewer was intimately aware of this procedure. He would thus interview one "correct" household and provide contact details. For that cluster that would be

the only questionnaire with a telephone number/contact detail. The other questionnaires probably all contained invented information. At the office, the control staff would phone the contactable household. That household would confirm the correctness of the interview. Since back checking was done only on a sample of households, the control staff then assumed that all the other interviews in the cluster were correct, although the other questionnaires were in fact falsifications.

About the only way to discover such a practice is by re-visiting all the selected households in a particular cluster.

Interviewing somebody else on behalf of the respondent

Proxy responses are permitted in many surveys. In the "Causes of migration" survey, proxy responses were allowed in the household section, but not in the case of information pertaining to a specific individual questionnaire. What was found during the quality control exercise? In a number of cases, the interviewer arrived at a household and completed the household section of the questionnaire. Then, using the grid, a respondent was randomly selected. However, upon finding that the selected person was not at home, the interviewer proceeded to interview somebody else in the household on behalf of the selected respondent. The interviewer should have waited for the selected respondent, or returned at another time or day. During the quality control re-visits, it took some time to ascertain what actually happened. On the surface it appeared as if the correct household was selected and interviewed. Members of the household would also confirm that the selected respondent is a member of that household. And if the person was not there during the revisit, there was a strong temptation to regard this as a correctly completed interview. But in a number of cases, when asked about the day the interviewer visited, it transpired that the interviewer did not actually interview the respondent. A family member was interviewed on behalf of him or her.

Interviewers followed this procedure for very much the same reasons they interviewed wrong households. They did not want to return, either due to an imposed tight schedule or for other reasons. The only way to expose this practice is to conduct a re-visit and spend time with the household enquiring about specific details of the interview, especially if the respondent is not at home.

Interviewing the wrong respondents

The purpose of using a grid to select respondents is to prevent severe bias in the results. If interviewers were allowed to choose respondents, the sample drawn would be a reflection of those persons mostly at home (more female and older). During the quality control visits it appeared that some interviewers were unable to use the grid (even after training). Consequently, their respondent selections were all wrong. This pointed to inadequate training and the selection of inappropriate interviewers.

Other interviewers were familiar with how to use the selection grid. Yet wrong respondents were selected for the individual interviews. Why did they do that? Invariably the selected respondent was not at home and instead of returning, one of the available family were interviewed without giving any attention to the possible consequences of such actions. Conducting a basic control/checking exercises at the office could have identified such problems early in the survey.

Even though much emphasis was placed on using a selection grid to select an unbiased sample of individual respondents, the results of the survey showed a bias towards female respondents.

Visiting the wrong households

During the series of revisits a number of cases were encountered where patently wrong households were selected. In certain enumerator areas, the selected dwellings/households were clearly indicated on the map. Yet the interviewers visited other households. In certain cases, the addresses of the households actually visited were on the questionnaires, while in other enumerator areas the address of the selected household was on the questionnaire, but the interview was not conducted there. In other cases again, the interviewers omitted to write an address on the questionnaire, clearly as a ploy to prevent anybody to ascertain afterwards where such an interview took place.

As an example: In one enumerator area the selected households were clearly shown on the map. Yet a revisit to those households showed that the interviewers did not visit them. By chance, it was established that one interview took place a number of blocks away from the selected enumerator area. When visiting that household, the researcher were informed of interviews in other households even further away. In considering the facts, it was our opinion that the interviewers did not complete the questionnaires in the wrong area because they were unable to read the map, but because the area selected in the sample was located in a lower socio-economic stratum, whereas the dwellings where the interviews were completed were located in a "better" area.

A number of other factors also contributed to interviews being done in a wrong area or at an incorrect household.

Bad quality maps/uncertain geography

In a small number of enumerator areas, genuine uncertainty existed about the location of and the specific boundaries of a specific enumerator area. Since the sample was drawn using census enumerator area information, the survey was dependent on maps depicting this area. In some areas no maps were available or the census office could only supply inadequate maps. This was a source for disagreement about the exact location and/or boundaries of the enumerator area. A small number of enumerator areas had to be replaced for this reason.

Inability of interviewers to read maps

In a number of enumerator areas the interviewers conducted interviews outside the designated enumerator area. One explanation was that some interviewers had difficulty in reading maps. However, as they worked in teams with a supervisor, this was not an excuse often given. Rather this occurred in combination with inaccurate maps.

Incomplete interviews

During the revisits, cases were found where, although an interview was conducted with the correct respondent, interviews were not completed in full. This seemed to occur in cases where the respondent or interviewer was in a hurry. The interviewer then only asked a selection of questions, skipping some. The length of the questionnaire may have contributed to this practice.

Wrong answers

During the revisits, the researchers came across cases where the answers provided during the re-interview did not match the answers provided during the original interview. Most commonly, these "errors" were related to attitude questions and those questions pertaining to migration intentions. Respondents changed their story with the lapse of time or when asked by a different person. However, factual errors were also noted. These applied specifically to a question whether a person had ever lived outside "this area" (see the discussion regarding this question in the main section of the report).

BIBLIOGRAPHY

Babbie, E.R. 1973. *Survey Research Methods*. Belmont: Wadsworth Publishing Company.

Bilsborrow, R.E., Oberai, A.S., Standing, G.1984. *Migration surveys in low income countries: Guidelines for survey and questionnaire design*. London: Croom Helm.

Bogue, D.J., Arriaga, E.E, Anderton, D.L. (Project Editors). 1993. *Readings in Population Research Methodology. Volume 4. Nuptiality, Migration, Household and Family Research*. Chicago: Social Development Center.

De Vaus, D.A. *Surveys in social research. Contemporary Social Research Series: Number 11*. London: George Allen & Unwin.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. 1953. *Sample survey methods and theory, 2 Vols*. New York: John Wiley & Sons, Inc.

Kish, L. 1965. *Survey sampling*. New York: John Wiley & Sons, Inc.

Kok, P.C., Wentzel, M.E and Pietersen, J. 2003. *Recent HSRC migration surveys, the data they have generated, methodological details, and some highlights*. Paper presented at the HSRC Workshop on migration, March 2003.

Kok, P.C., O'Donovan, M, Bouare, O and Van Zyl, J.A. 2002. *Post-apartheid patterns of internal migration in South Africa*. Pretoria: HSRC.

Kok, P.C., O'Donovan. 2001. *The causes of internal migration in South Africa: Findings from a preliminary survey*. Pretoria: HSRC. (Unpublished report).

Levy, P.S., Lemeshow, S. 1999. *Sampling of populations: Methods and Applications*. Third edition New York: Wiley and Sons Inc.

Marsh, C. 1982. *The survey method: The contribution of surveys to sociological understanding. Contemporary Social Research Series: Number 6*. London: George Allen & Unwin.

Shisana, O., Simbayi, L. 2002. Nelson Mandela/HSRC study of HIV/AIDS: *South African national HIV prevalence, behavioural risk and mass media: Household survey 2002.* Pretoria: HSRC.

Shryock, H.S. and Siegel, J.S. 1976. *The Methods and Materials of Demography.* Condensed edition by Stockwell, E.G. New York: Academic Press.

United Nations. 1971. *Methodology of Demographic Sample Surveys.* Report of the Inter-regional Workshop on Methodology of Demographic Sample Surveys. New York: United Nations. Statistical Papers Series M. No. 51.

United Nations. 1982. *National Household Survey Capability Programme. Non-sampling errors in household surveys: Sources, Assessment and Control.* New York: United Nations.

United Nations. 2000. *Report on the workshop on application of new information technology to population data.* Bangkok, October 1999. Economic and Social Commission for Asia and the Pacific.

Warwick, D.P., Lininger, C.A. 1975 *The sample survey: Theory and practice.* New York: McGraw-Hill Book Company