Statistics South Africa

CENSUS 2001

Post-enumeration survey

Results and methodology

Report no. 03-02-17 (2001)

Statistics South Africa
2004

Pali Lehohla
Statistician-General

Census 2001: Post-enumeration survey: Results and methodology

Published by Statistics South Africa, Private Bag X44, Pretoria 0001

© Statistics South Africa 2004
Users may apply or process this data, provided Statistics South Africa (Stats SA) is acknowledged as the original source of the data; that it is specified that the application and/or analysis is the result of the user's independent processing of the data; and that neither the basic data nor any reprocessed version or application thereof may be sold or otherwise offered for sale in any form whatsoever without prior permission from Stats SA.

ISBN: 0-621-34333-1

Stats SA Library Cataloguing-in-Publication (CIP) Data
Census 2001: Post-enumeration survey: Results and methodology / Statistics South Africa.
Pretoria: Statistics South Africa, 2004
p. 89 [Report No. 03-02-17(2001)]
ISBN: 0-621-34333-1
1. Census
2 Census undercounts
3 Census – methodology

A complete set of Stats SA publications is available at Stats SA Library and the following libraries:

National Library of South Africa, Pretoria Division
National Library of South Africa, Cape Town Division
Library of Parliament, Cape Town
Bloemfontein Public Library
Natal Society Library, Pietermaritzburg
Johannesburg Public Library
Eastern Cape Library Services, King William's Town
Central Regional Library, Polokwane
Central Reference Library, Nelspruit
Central Reference Collection, Kimberley
Central Reference Library, Mmabatho

This report is available on the Stats SA website: *www.statssa.gov.za*

Copies are available from: Printing and Distribution, Statistics South Africa
Tel: (012) 310 8251
(012) 310 8161
Fax: (012) 322 3374
(012) 310 8619
E-mail:distribution@statssa.gov.za

Please also consult the website for a full list of printed and electronic census products, both already available and forthcoming. Most products can be ordered as detailed as above or downloaded from the website.

Prologue

A 'perfect' census is impossible, but census figures are still valuable if the quality and the limitations of the data are understood by the users. An assessment of the magnitude and direction of the errors in a census is thus necessary to respond to questions about the results and to attacks on their accuracy.

As part of the quality check for Census 2001, a Post-Enumeration Survey (PES) was conducted in November 2001, approximately one month after the census. Fieldworkers re-visited a scientifically selected sample of almost 1% of the census enumeration areas, to do an independent recount. The published census results are adjusted for undercount according to the findings of the PES. In addition to the check on coverage, the PES also involved an independent re-measurement of the basic characteristics of the population.

This report has been prepared in two parts: Part I – Results and Analysis, to present and explain the PES results, and Part II – Methodology and Procedural History, to acquaint data users with the methods used and to provide an account of the procedures as they were implemented. Relevant PES definitions are provided in Appendix I.

This report does not cover all the details of the design and implementation of the PES, since those are numerous and beyond the scope of this document. In the context of the PES, many internal documents were developed, such as the Fieldworker's Manual, the Matching Manual, and others. Interested users may contact Stats SA to obtain these documents.

Acknowledgements

The Post-Enumeration Survey was accomplished through the diligent efforts of many people.

Statistics South Africa

Stats SA conducted the PES and was responsible for all its developmental and implementation aspects. Various teams contributed to the PES. The census sub-projects staff developed the Geographic Information System and the census maps used in the PES. The Quality and Methodology staff provided managerial oversight and ensured the independence of the PES from the census organisational units. The PES sub-project staff planned, developed, and managed all aspects of the PES. The Household Surveys staff was responsible for planning the training and the field operations while the provincial survey staff provided the training and field coordination. The fieldworkers and office workers were the 'foot soldiers' who carried out the field visits and processed the questionnaires in the office. The Publishing staff reviewed and printed manuals and survey materials.

Statistics Council

The Council, in particular the Census Subcommittee, provided technical review and feedback to help make the PES results meaningful.

US Census Bureau

The United States Census Bureau provided technical assistance in all aspects of the Post-Enumeration Survey, including questionnaire design, development of manuals and training guides, methodology, procedural specifications, and analysis. This assistance was funded by the United States Agency for International Development (USAID).

Contents

List of tables

List of figures

Part I

RESULTS AND ANALYSIS

1. INTRODUCTION

1.1 Objectives of the post-enumeration survey

A post-enumeration survey (PES) is a sample survey conducted immediately after a census, for the primary purpose of evaluating the census. It provides a concrete statistical basis for estimating census coverage, that is, the extent of undercount or overcount, and for adjusting the census data if there is evidence of significant coverage error. It also provides an evaluation of the reliability of some of the characteristics reported in the census.

There are many reasons why people might not be included in a census count:

- failure to account for all inhabited areas in the frame of census enumeration areas (EAs)
- boundary demarcation problems or boundary interpretation problems causing overlap or omission of parts of EAs
- incomplete listing of dwellings within EAs (failure to identify all places where people might live)
- failure to visit all listed dwellings
- failure to identify all households, where multiple households exist within dwellings
- failure to obtain interviews for all households (non-contact, refusals, non-return of questionnaires left for self-enumeration)
- failure to identify all persons within households
- incomplete or poor-quality information on persons for key variables
- failure to observe the association (inclusion) rule which, in Census 2001, is based on the presence of the individual in the household on census reference night (*de facto* coverage definition)
- lost or unprocessable questionnaires

In Census 2001, underenumeration and overenumeration errors were corrected through the coverage adjustment process and the results were thus made more accurate and complete. For groups that were more seriously undercounted (or overcounted) than others, the adjustment prevented their distortion in the distribution of totals in the population.

In the PES, a scientifically selected sample of the target census population is independently re-enumerated. Coverage status is determined through case-by-case comparison of the new enumerated cases with the original census records (see Section 7 on matching). This two-way match and a field follow-up exercise allow the identification of omissions as well as erroneous inclusions, and estimates of population totals are based on direct observation. In non-matching studies, conclusions are based upon indirect evidence, and sometimes on the judgment of the analyst.

The PES methodology is discussed in detail in Part II of this report.

In addition to coverage adjustments, the PES:

- assists data users in using the census data judiciously by giving them greater insight into the quality and the limitations of the data;
- assists in providing a better basis for demographic projections;
- helps evaluate the quality of the enumeration areas and maps as a sampling frame for intercensal household surveys; and

- permits an assessment of the effectiveness of the census design, management, and procedures in order to improve the planning and implementation of future censuses.

1.2 PES target universe

PES 2001 sought to estimate the total number of persons and households in housing units and workers' hostels[1] on the night of 9-10 October 2001 (census night). The units of observation were the persons who spent the census night and/or the PES night in these living quarters.

The source for the PES sampling frame was the database of enumeration areas (EAs) demarcated for the census (Section 6.1). While all four geography types – namely, urban formal, urban informal, tribal areas, and rural formal – were within the scope of coverage of the PES, all EA types were not. The sampling frame was restricted in scope to certain EA types: farm, hostel, informal settlement, smallholding, tribal settlement, and urban settlement. Industrial, institutional, recreational, and vacant EA types were out-of-scope.

In terms of living quarters, the PES universe includes only:
- persons living in non-seasonal housing units
- persons living in hostels for workers

The balance of the population (approximately 5% of the total), consisting of those living in other types of collective quarters or those in no living quarters, is excluded from the scope of the PES because special methodologies would be required. Hence, the PES does not represent people in:
- residential hotels
- homes for the aged
- student residences
- tourist hotels/motels/inns, and
- institutions

or the homeless on the street.

2 COVERAGE EVALUATION OF CENSUS 2001 – PERSONS

2.1 Estimation of 'true population'

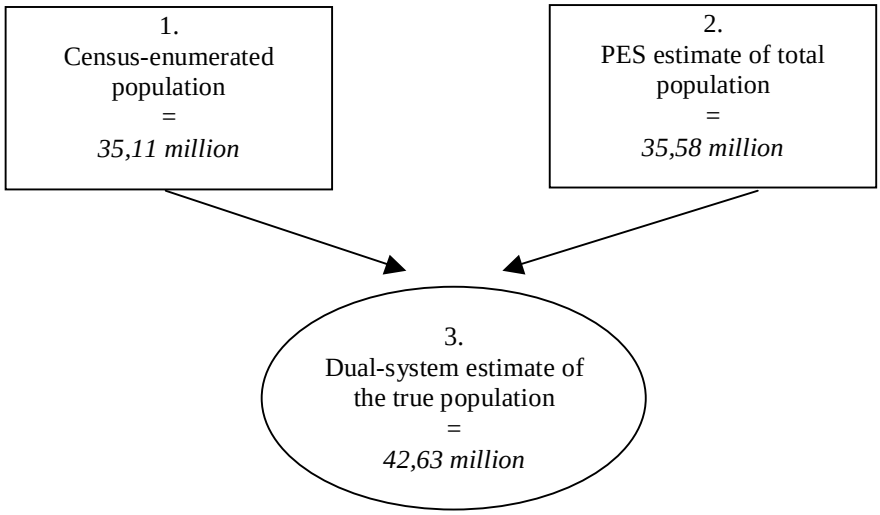
Two independent sources or 'systems' are used to arrive at the estimate of the *true population*: the census and the PES. The first attempt at measuring the true population yields the *census-enumerated population*, based on an exhaustive enumeration. The second attempt yields the *PES estimate of the total population*, based on sampling techniques.

Instead of assuming that one or the other is better, both of these estimates are used to derive a third, composite estimate of the true population called the 'dual-system estimate of the true population' (see Section 8.2 for estimation formulas). The dual system provides an estimate of the cases included in one source (PES) and excluded from the other (Census), and vice versa. Both estimates contribute to the dual-system estimate, which is more complete than either the census or the PES estimate alone.

[1] Note that 'hostel' refers to workers' hostel throughout, and does not include student residences or boarding school hostels.

In the end, this *true population* is compared with the *census-enumerated population* and the difference is the net *undercount* (or *overcount*).

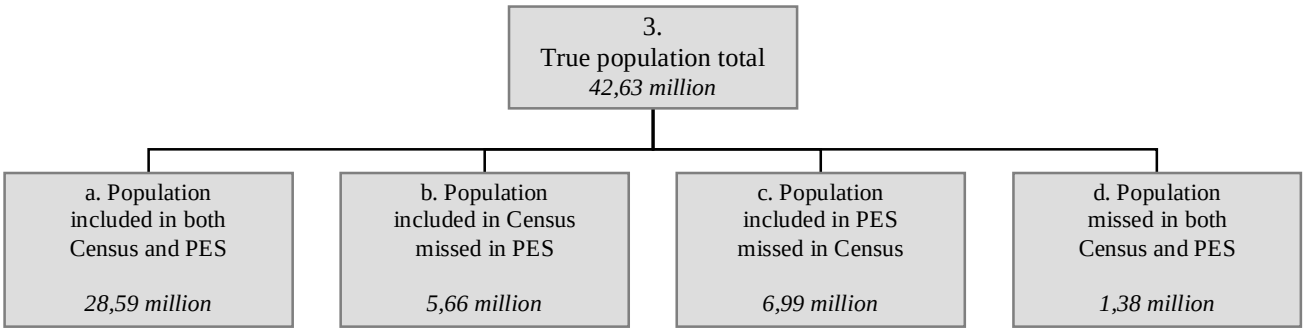
Figure 2.1
Estimates of total population from the individual systems and from the dual system – in-scope subuniverse



In the in-scope subuniverse, the separate census and PES enumerations produced 35,11 million and 35,58 million persons, respectively. Using the dual-system estimation method, the true population of South Africa in the in-scope subuniverse was estimated at 42,63 million.

Four components together make up the dual-system estimate of the true population.

Figure 2.2
Breakdown of dual-system estimate of population total – in-scope subuniverse



Note: Sums are subject to rounding error.

Components (a), (b), and (c) are obtained through a **matching** process, based on direct observation. Component (d) is a mathematical derivation, based on an assumption of independence.

Component (a), the population included in both the census and the PES, was estimated at 28,59 million persons; component (b), the population included in the census but missed in the PES, was estimated at 5,66 million; component (c), the population included in the PES but missed in the census, was estimated at 6,99 million; and component (d), the population missed in both the census and the PES, was estimated at 1,38 million (derivations can be found in Figure 8.4).

In Table 2.1, it can be seen that, of the 35,11 million persons counted in the census for the in-scope subuniverse, 34,25 million are estimated to be correctly enumerated. Of these, the PES included 28,59 million and missed 5,66 million. The census erroneous inclusions (fabrications, duplications, and geographic misallocations) are estimated to be 0,86 million or approximately 2,4% of the census total.

Table 2.1
Coverage distribution of Census population – in-scope subuniverse
(in millions rounded to two decimals)

	Census enumeration
Total excluding erroneous inclusions	34,25
Included in PES	28,59
Omitted from PES	5,66
Erroneous inclusions	0,86
Total including erroneous inclusions	**35,11**

It is estimated that the census omitted 8,37 million persons in total, 6,99 million of which were correctly enumerated in the PES and another 1,38 million of which were missed in the PES as well as the census (Table 2.2). This total omission does not take into account what it added incorrectly (the erroneous inclusions). When it is offset by the 0,86 million erroneous inclusions, the net undercount is 7,51 million. The net undercount relative to the 42,63 million in the true population is thus 17,6% (Table 2.5).

Table 2.2
Coverage distribution of true population – in-scope subuniverse
(in millions rounded to two decimals)

		Census enumeration		
		Included	Omitted	**Total**
PES Population	Included	28,59	6,99	**35,58**
	Omitted	5,66	1,38	**7,04**
	Total excluding erroneous inclusions	**34,25**	**8,37**	**42,63**

Sums are subject to rounding error.

While the PES estimated the total population in the in-scope subuniverse at 35,58 million, it omitted 5,66 million persons who were correctly enumerated in the census, and another 1,38 million who were missed in both the census and the PES, for a total omission of 7,04 million (Table 2.2).

The total South African population of 44,8 million persons was calculated by adding the census-enumerated 2,2 million persons in the other collective living quarters and the out-of-scope EA types to the dual-system estimate of 42,6 million in the in-scope subuniverse (Table 2.3).

Table 2.3
Unadjusted and adjusted Census population – full universe
(in millions rounded to two decimals)

	Persons in housing units and hostels within in-scope EA types	**Persons in other collective living quarters and other EA types**	**Total population**
Unadjusted	35,11	2,19	**37,30**
Adjusted	42,63	2,19	**44,82**

The overall empirical probabilities of inclusion and omission of a person in the census or in the PES are shown below in Table 2.4 (derivations can be found in Figure 8.3). According to the enumeration results, a member of the in-scope subuniverse had approximately an 80,4% chance of being enumerated in the census, an 83,5% chance of being enumerated in the PES, and a 67,1% chance of being enumerated in both. Conversely, the person had approximately a 13,3% chance of being included in the census but missed in the PES, a 16,4% chance of being included in the PES but missed in the census, and a 3,3% chance of being missed in both.

Table 2.4
Probabilities of inclusion and omission of a person – in-scope subuniverse

Probability of being included in the census	0,8036
Probability of being included in the PES	0,8348
Probability of being included in both the census and the PES	0,6708
Probability of being included in census, but missed in the PES	0,1328
Probability of being included in the PES, but missed in the census	0,1640
Probability of being missed in both the census and the PES	0,0325

2.2 Estimation of the net undercount rate

The net undercount (or overcount) is the difference between the estimated true population (dual-system estimate) and the census-enumerated population. The rate is the net undercount expressed as a percentage of the estimated true population.

The net undercount rates, together with their absolute errors and confidence intervals, are shown in the following tables for geographic[1] and demographic groups. The confidence interval is formed around the estimate by adding or subtracting the absolute error. It must be noted that high absolute errors indicate that the estimate is not statistically reliable and confidence intervals are very wide as a result.

In Table 2.5, it can be observed that the net undercount rate at the national level was estimated at 17,6%, with possible values ranging from 16,6% to 18,7%.

When comparing rates for different sets of persons, the confidence intervals must be taken into account. Before concluding that a 'differential' undercount exists, for example, that the undercount rate for one group is in fact higher (or lower) than that of another group, the two confidence intervals in question must not overlap. (This is equivalent to a two-tailed hypothesis test at the 0,05 level of significance.) An overlap in the intervals indicates that – except for a 5% chance of erring in the conclusion – the difference observed is not statistically significant due to random error, in other words, that there is no evidence of a real difference. A 'floating bars' chart is useful for visualising the intervals (see Figure 2.3a-f).

Table 2.5
Net undercount rate for persons by province –
in-scope subuniverse
(values expressed in percentage points rounded to one decimal)

			95% Confidence interval limits*	
Category	**Net undercount rate**	**Absolute error (+ or -)**	**Lower**	**Upper**
All persons	17,6	1,1	16,6	18,7
Province				
Eastern Cape	14,7	1,8	12,9	16,6
Free State	17,6	1,2	16,4	18,9
Gauteng	18,7	3,5	15,3	22,2
KwaZulu-Natal	22,5	5,6	16,9	28,1
Limpopo	14,4	0,4	14,0	14,7
Mpumalanga	16,1	1,0	15,1	17,0
North West	16,0	1,2	14,8	17,3
Northern Cape	14,1	0,8	13,2	14,9
Western Cape	16,3	1,5	14,8	17,7

* subject to rounding error

Among the provinces, the highest undercount was observed in KwaZulu-Natal, Gauteng and Free State (22,5%, 18,7%, and 17,6%, respectively) (see Table 2.5 and Figure 2.3a).

[1] PES tables for urban/non-urban splits will be produced only after the revised definition is finalised by Stats SA.

However, due to overlap in the confidence intervals as seen in the figure, not all differences are significant. There is no significant undercount difference among KwaZulu-Natal, Gauteng, and Free State. Undercounts in the provinces of Eastern Cape, Free State, Gauteng, Mpumalanga, North West, and Western Cape are not significantly different from one another. The undercount in KwaZulu-Natal is significantly higher than in Eastern Cape, Limpopo and Northern Cape, but not significantly higher than in the other provinces. The lowest observed undercount, in Northern Cape, is significantly lower than that in Free State, Gauteng, KwaZulu-Natal and Mpumalanga, but not significantly lower than in the other provinces.

Figure 2.3a
Graphic representation of confidence intervals for persons undercount rate – provinces

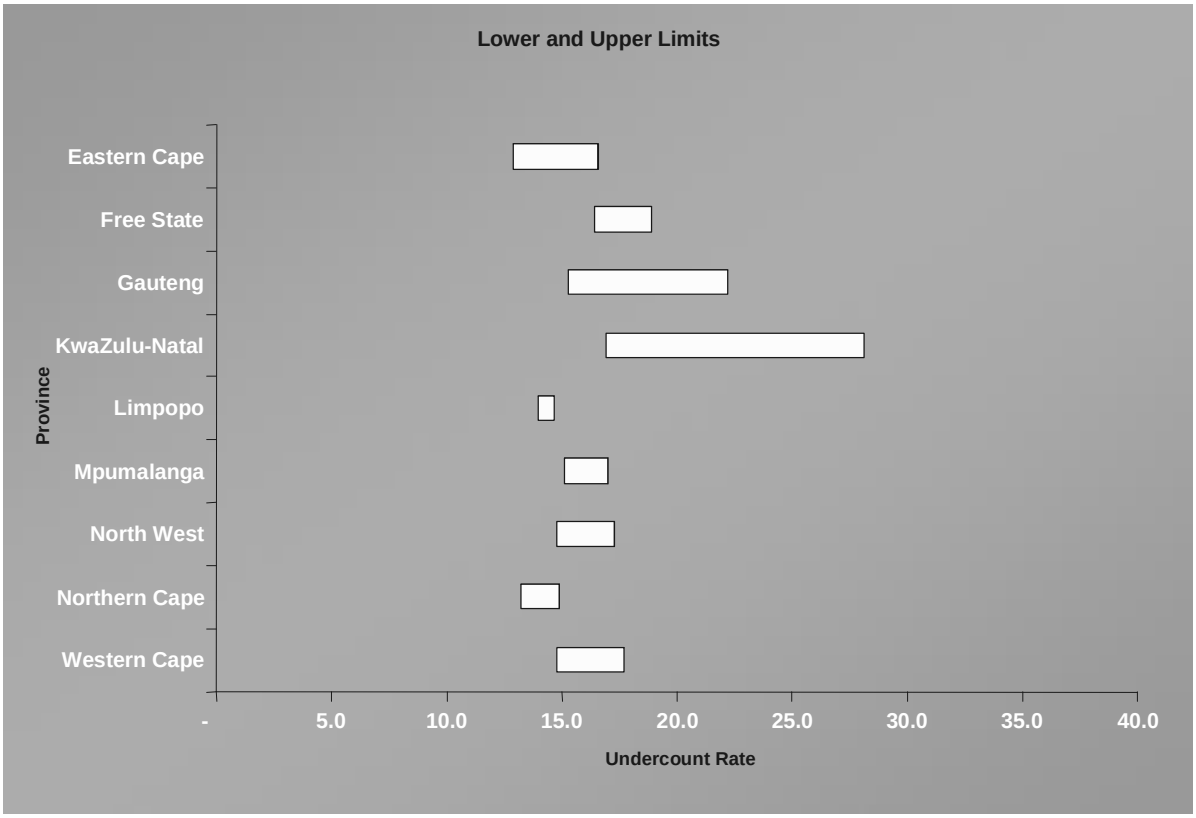


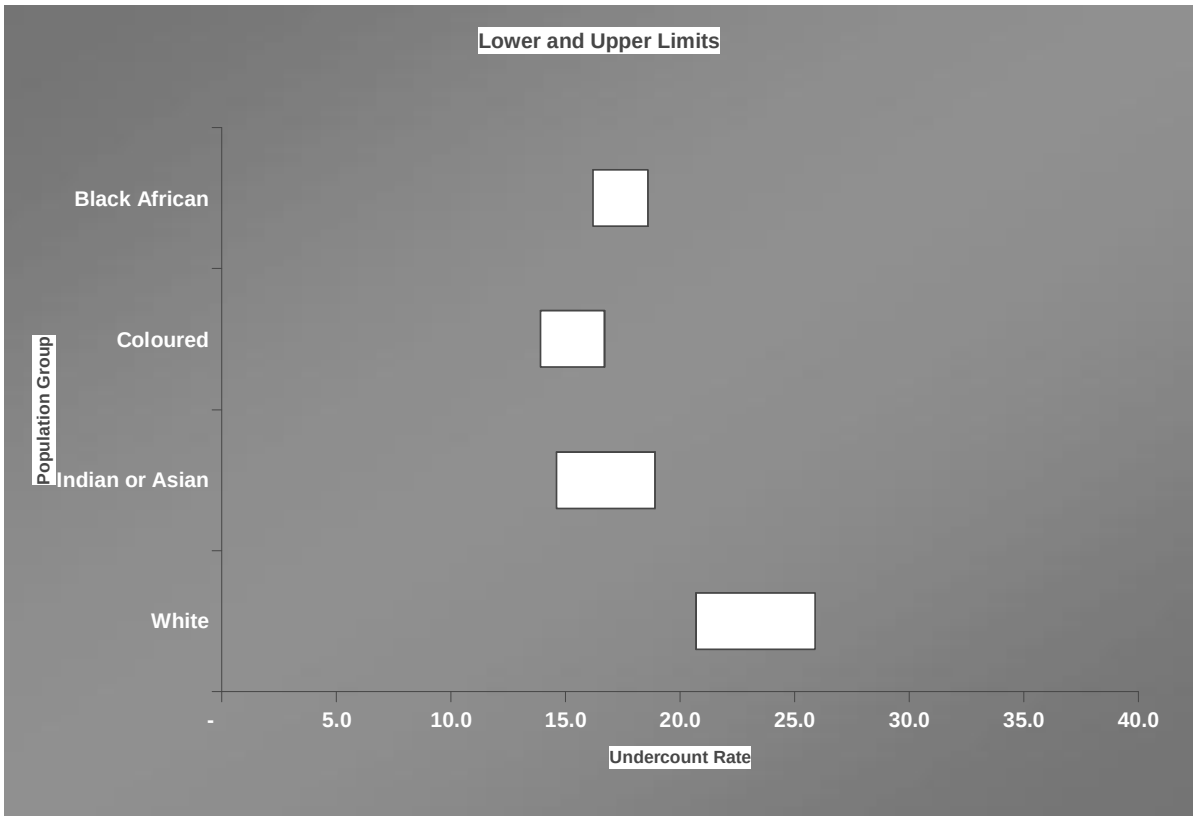
Table 2.6a
Net undercount rate for persons by demographic group – in-scope subuniverse
Single-variable classifications
(values expressed in percentage points rounded to one decimal)

			95% Confidence interval limits*	
	Net undercount rate	**Absolute error (+ or -)**	**Lower**	**Upper**
All persons	17,6	1,1	16,6	18,7
Population group				
Black African	17,4	1,2	16,2	18,6
Coloured	15,3	1,4	13,9	16,7
Indian or Asian	16,7	2,1	14,6	18,9
White	23,3	2,6	20,7	25,9
Sex				
Male	18,6	1,0	17,6	19,7
Female	16,9	1,1	15,8	18,0
Age group				
Under 5 years	16,8	1,2	15,6	18,1
5-9 years	16,2	1,4	14,8	17,6
10-14 years	16,0	1,5	14,5	17,5
15-19 years	16,6	1,1	15,4	17,7
20-29 years	20,6	1,0	19,7	21,6
30-44 years	19,6	1,0	18,6	20,6
45-64 years	16,8	1,0	15,8	17,8
65+ years	14,6	1,2	13,5	15,8

* subject to rounding error

At first glance, the undercount for males (18,6%) seems higher than that for females (16,9%). However, an inspection of the intervals (17,6–19,7% vs. 15,8–18,0%) reveals that this observed difference is not statistically significant. There is thus insufficient evidence to conclude that a differential undercount between males and females occurred in reality. When population groups are compared, the highest undercount is found among whites (23,3%) (see Figure 2.3b below). Whites were undercounted at a significantly higher rate than the other population groups while there is no significant difference in undercount among the African, coloured, and Indian/Asian groups.

Figure 2.3b
Graphic representation of confidence intervals for persons undercount rate – population groups



Except for the 20-29 years group and the 30-44 years group, no claim of a differential undercount among age groups can be made, that is, the undercount rate is in the same range for all the other age groups (see Figure 2.3c below). While these two age groups are significantly more undercounted than the other groups, their undercounts are not significantly different from each other.

Figure 2.3c
Graphic representation of confidence intervals for persons undercount rate: age groups

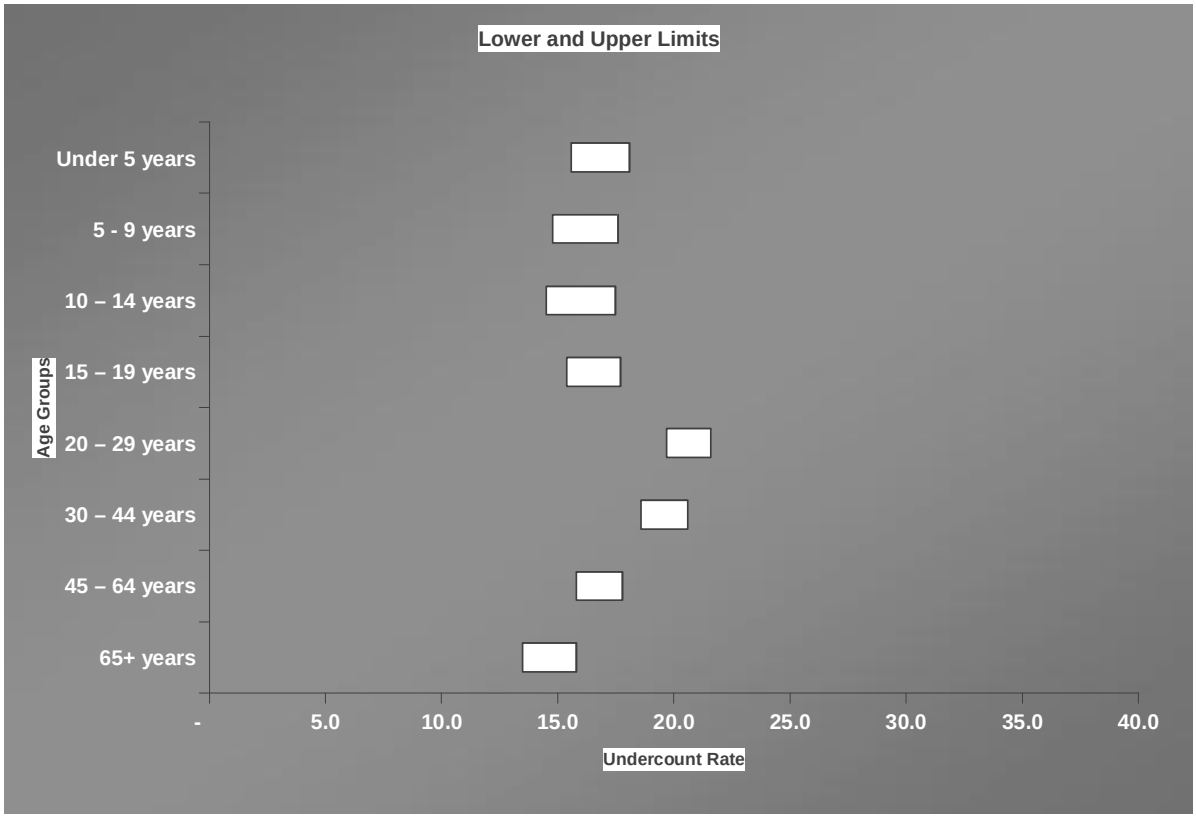


Table 2.6b
Net undercount rate for persons by demographic group – in-scope subuniverse
Two-variable classifications
(values expressed in percentage points rounded to one decimal)

				95% Confidence interval limits*	
		Net undercount rate	Absolute error (+ or -)	Lower	Upper
Population group by sex					
Black African	Male	18,5	1,2	17,3	19,7
Black African	Female	16,6	1,3	15,3	17,8
Coloured	Male	15,7	1,5	14,2	17,2
Coloured	Female	15,2	1,3	13,8	16,5
Indian/Asian	Male	17,2	2,1	15,0	19,3
Indian/Asian	Female	16,5	2,2	14,3	18,7
White	Male	23,9	2,3	21,6	26,1
White	Female	23,1	3,0	20,1	26,1
Population group by age					
Black African	Under 5 years	16,9	1,4	15,5	18,3
Black African	5-9 years	15,9	1,6	14,3	17,5
Black African	10-14 years	15,6	1,7	14,0	17,3
Black African	15-19 years	16,2	1,3	14,9	17,5
Black African	20-29 years	20,9	1,2	19,8	22,1
Black African	30-44 years	19,1	1,2	18,0	20,3
Black African	45-64 years	16,1	1,2	14,9	17,3
Black African	65+ years	13,9	1,3	12,6	15,2
Coloured	Under 5 years	15,8	1,6	14,2	17,4

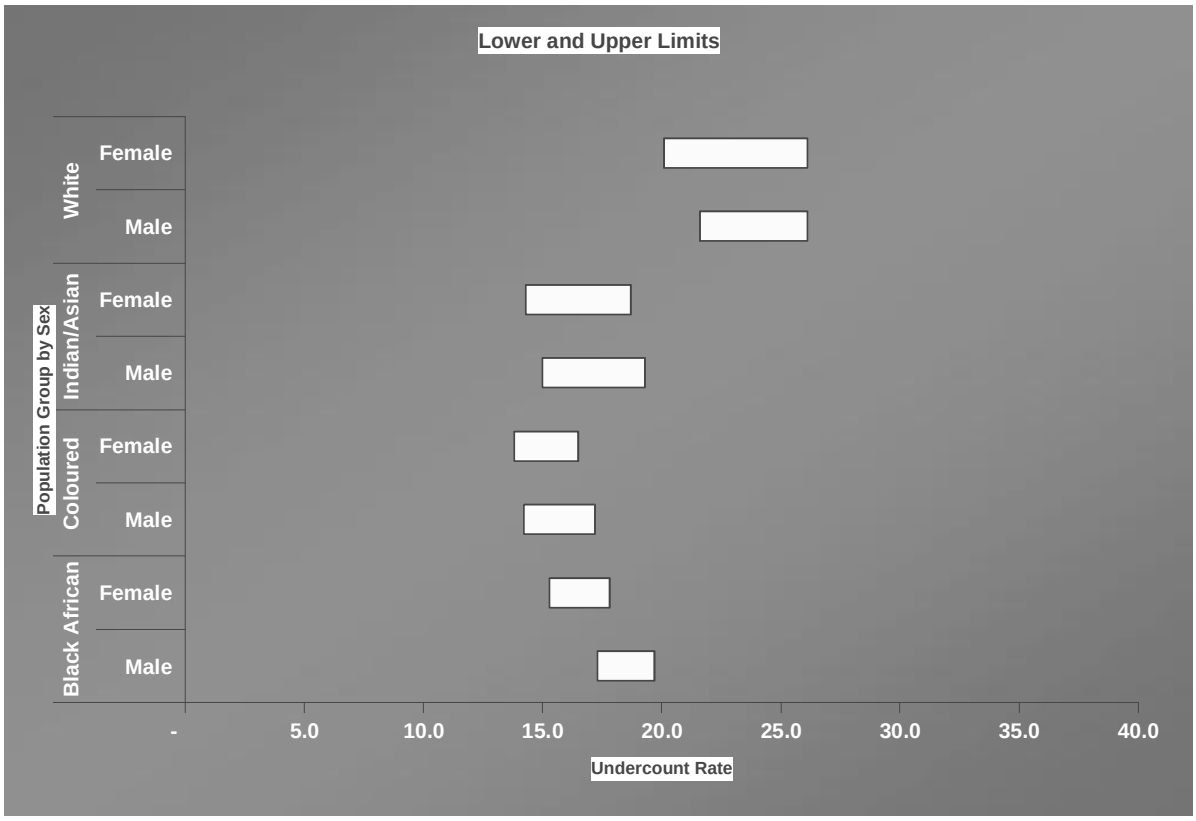
				95% Confidence interval limits*	
		Net undercount rate	Absolute error (+ or -)	Lower	Upper
Coloured	5-9 years	15,2	1,4	13,8	16,7
Coloured	10-14 years	14,9	1,6	13,3	16,4
Coloured	15-19 years	15,3	1,3	13,9	16,6
Coloured	20-29 years	16,9	1,0	15,8	17,9
Coloured	30-44 years	16,0	1,4	14,6	17,4
Coloured	45-64 years	13,8	1,9	12,0	15,7
Coloured	65+ years	13,3	3,2	10,2	16,5
Indian/Asian	Under 5 years	15,5	3,5	12,0	18,9
Indian/Asian	5-9 years	14,6	3,2	11,5	17,8
Indian/Asian	10-14 years	14,7	1,4	13,3	16,2
Indian/Asian	15-19 years	15,4	2,1	13,3	17,5
Indian/Asian	20-29 years	17,8	1,6	16,2	19,5
Indian/Asian	30-44 years	18,9	2,2	16,7	21,1
Indian/Asian	45-64 years	16,8	2,7	14,2	19,5
Indian/Asian	65+ years	15,2	4,8	10,4	20,0
White	Under 5 years	21,3	3,6	17,7	24,9
White	5-9 years	25,1	5,0	20,1	30,1
White	10-14 years	25,0	4,1	20,9	29,1
White	15-19 years	24,8	3,3	21,5	28,1
White	20-29 years	24,1	1,7	22,4	25,8
White	30-44 years	26,9	2,5	24,4	29,3
White	45-64 years	21,5	2,9	18,7	24,4
White	65+ years	18,1	2,7	15,4	20,8
Sex by age					
Male	Under 5 years	17,0	1,3	15,7	18,3
Male	5-9 years	16,4	1,5	14,9	17,8
Male	10-14 years	16,2	1,5	14,7	17,7
Male	15-19 years	16,7	1,1	15,6	17,8
Male	20-29 years	22,3	1,0	21,3	23,3
Male	30-44 years	21,5	1,0	20,6	22,5
Male	45-64 years	18,2	0,9	17,2	19,1
Male	65+ years	15,9	1,1	14,8	17,0
Female	Under 5 years	16,9	1,1	15,7	18,0
Female	5-9 years	16,2	1,4	14,8	17,6
Female	10-14 years	16,0	1,4	14,5	17,4
Female	15-19 years	16,6	1,2	15,4	17,9
Female	20-29 years	19,3	1,0	18,3	20,3
Female	30-44 years	18,1	1,1	17,0	19,2
Female	45-64 years	15,7	1,1	14,6	16,8
Female	65+ years	14,0	1,3	12,7	15,3

* subject to rounding error

Table 2.6b and the Figures 2.3d through 2.3f will allow the reader to make many different comparisons based on population group by sex, population group by age, or sex by age.

For example, the lack of a differential undercount between males and females holds across all population groups (Figure 2.3d). The same figure also shows that the undercount for both white males and white females is significantly higher than that for all other population groups by sex. White males are significantly more undercounted than all other population group/sex groups, except white females. Finally, except for the difference between black males and coloured females as well as black males and coloured males, there are no other significant differences in undercount among the population group/sex groups.

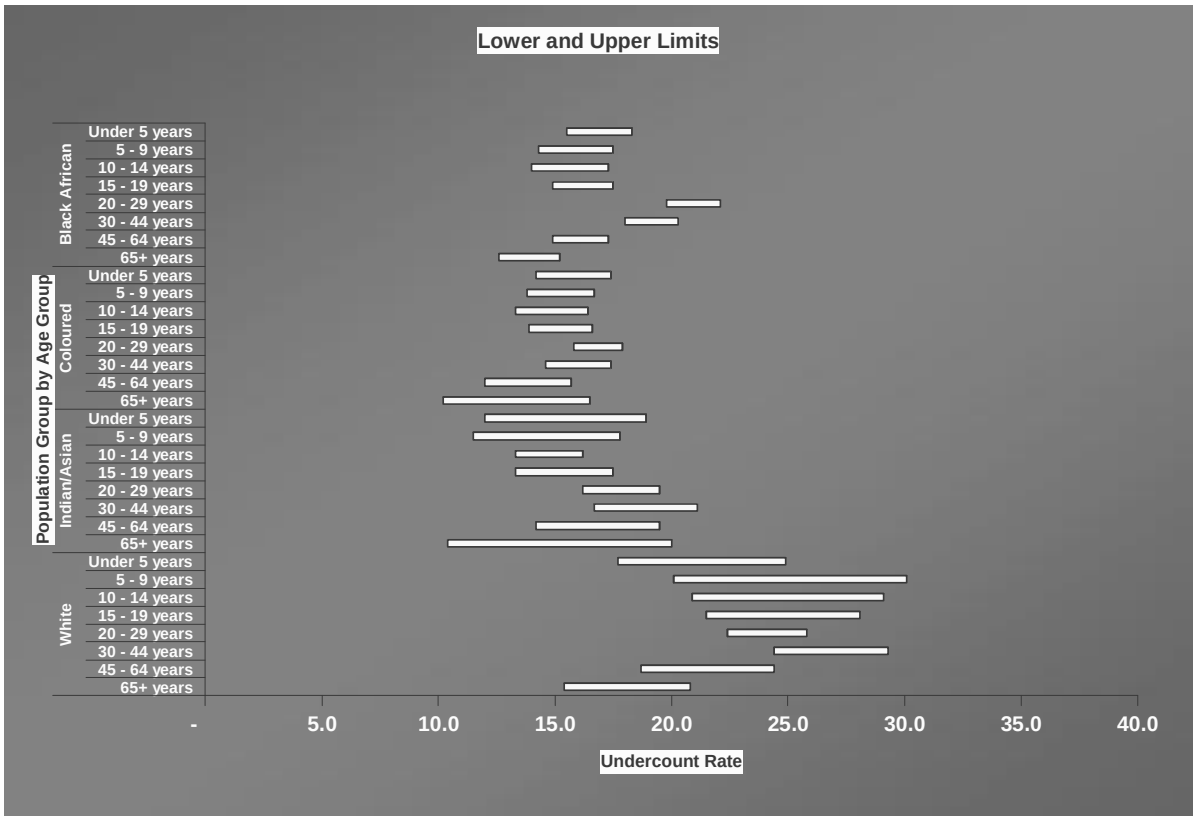
Figure 2.3d
Graphic representation of confidence intervals for persons undercount rate – population group by sex



Other interesting findings (Figure 2.3e) are, for example, that the undercount for whites 20-29 years old (22,4–25,8%) is significantly higher than that for Indians or Asians 20-29 years (16,2–19,5%), coloureds 20-29 years (15,8–17,9%) and black Africans 20-29 years (19,8%–22,1%).

Likewise, the undercount for males 20–29 years (21,3–23,3%) is significantly higher than that for females 20–29 years (18,3–20,3%).

Figure 2.3e
Graphic representation of confidence intervals for persons undercount rate – population group by age group



Another example (Figure 2.3f) is that, within both males and females, the undercounts for the age groups 20-29 years and 30-44 years are not significantly different from each other but they are both significantly higher than for the other age groups. However, where population groups are concerned (Figure 2.3e), the finding that the age group 20-29 years has a significantly higher undercount than all other age groups (except 30-44 years) holds only for black Africans. For coloureds, we are able to conclude that there is a differential undercount by age group only between the age groups 20-29 years (15,8–17,9%) and 45-64 years (12,0–15,7%). For the Indian/Asian group, the only evident differential undercount by age group is between 10-14 years (13,3% –16,2%) and 30-44 years (16,7% – 21,1%) and between 10-14 years (13,3–16,2%, actually 13,30–16,15%) and 20-29 years (16,2%–19,5%, actually 16,21–19,49%). For the white group, the only differential undercount by age in evidence is in the 65+ age group, where it is significantly lower than that for 10-14 years, 15-19 years, 20-29 years and 30-44 years.

Figure 2.3f
Graphic representation of confidence intervals for persons undercount rate – sex by age group

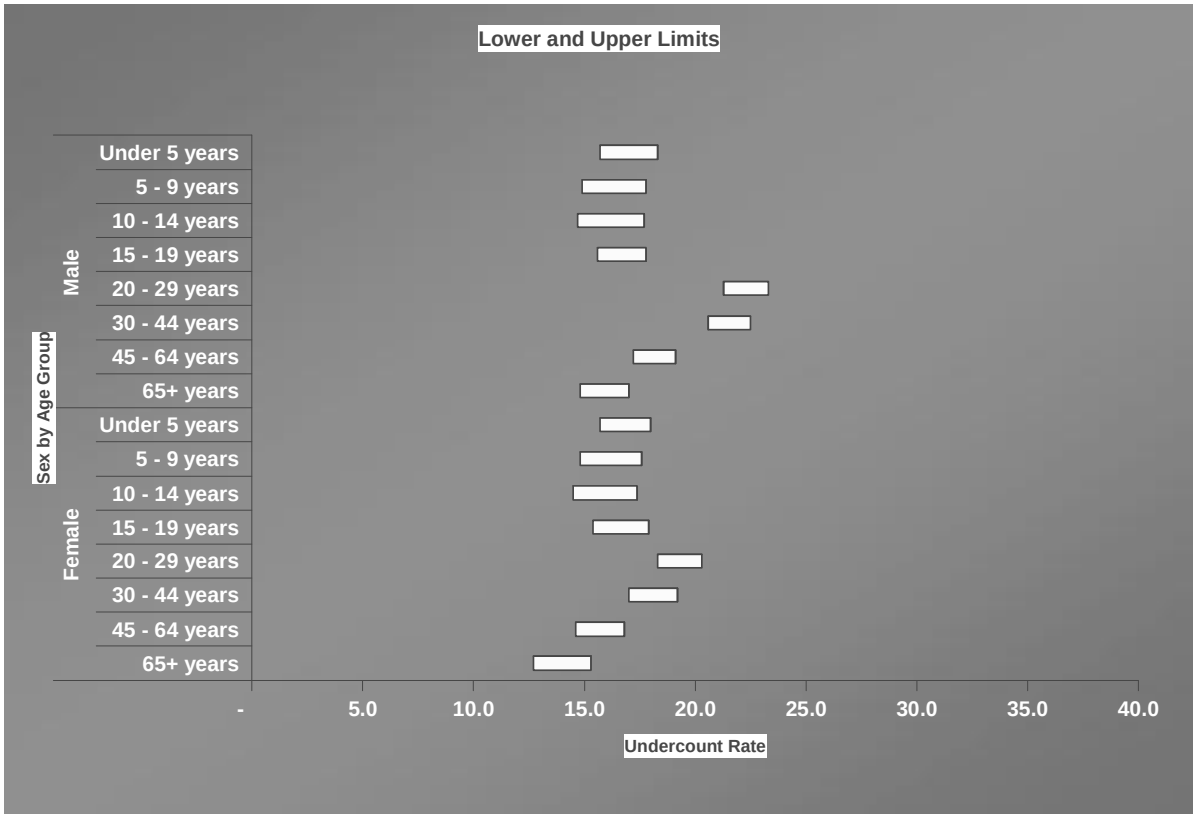


Table 2.6c
Net undercount rate for persons by demographic group – in-scope subuniverse
Three-variable classifications
(values expressed in percentage points rounded to one decimal)

					95% Confidence interval limits*	
			Net undercount rate	Absolute error (+ or -)	Lower	Upper
Population grp	**Sex**	**Age**				
Black African	Male	0-4	17,1	1,5	15,6	18,6
Black African	Male	5-9	16,1	1,7	14,4	17,7
Black African	Male	10-14	15,8	1,7	14,1	17,5
Black African	Male	15-19	16,3	1,3	15,0	17,5
Black African	Male	20-29	22,7	1,2	21,6	23,9
Black African	Male	30-44	21,5	1,1	20,4	22,6
Black African	Male	45-64	17,9	1,1	16,8	19,0
Black African	Male	65+	15,3	1,3	14,0	16,5
Black African	Female	0-4	17,0	1,3	15,7	18,2
Black African	Female	5-9	15,9	1,6	14,3	17,5
Black African	Female	10-14	15,6	1,6	14,0	17,2
Black African	Female	15-19	16,2	1,4	14,8	17,6
Black African	Female	20-29	19,4	1,2	18,2	20,6
Black African	Female	30-44	17,2	1,3	16,0	18,5
Black African	Female	45-64	14,8	1,3	13,5	16,1
Black African	Female	65+	13,3	1,4	11,9	14,7
Coloured	Male	0-4	15,9	2,2	13,7	18,1
Coloured	Male	5- 9	15,3	1,4	13,9	16,7
Coloured	Male	10-14	15,0	2,0	13,0	17,0
Coloured	Male	15-19	15,4	1,8	13,6	17,2
Coloured	Male	20-29	17,4	1,1	16,3	18,5

					95% Confidence interval limits*	
			Net undercount rate	Absolute error (+ or -)	Lower	Upper
Coloured	Male	30-44	16,6	1,4	15,2	17,9
Coloured	Male	45-64	14,3	1,8	12,5	16,1
Coloured	Male	65+	13,9	3,4	10,4	17,3
Coloured	Female	0-4	15,9	1,2	14,8	17,1
Coloured	Female	5-9	15,3	1,5	13,8	16,8
Coloured	Female	10-14	14,9	1,3	13,7	16,2
Coloured	Female	15-19	15,3	1,1	14,2	16,4
Coloured	Female	20-29	16,5	1,1	15,4	17,6
Coloured	Female	30-44	15,6	1,5	14,2	17,1
Coloured	Female	45-64	13,5	1,9	11,7	15,4
Coloured	Female	65+	13,0	3,3	9,8	16,3
Indian or Asian	Male	0-4	15,6	2,5	13,1	18,0
Indian or Asian	Male	5-9	14,7	4,2	10,5	18,9
Indian or Asian	Male	10-14	14,8	2,3	12,5	17,1
Indian or Asian	Male	15-19	15,5	3,1	12,4	18,6
Indian or Asian	Male	20-29	18,9	1,8	17,1	20,7
Indian or Asian	Male	30-44	19,3	2,0	17,4	21,3
Indian or Asian	Male	45-64	17,3	2,5	14,8	19,7
Indian or Asian	Male	65+	15,7	2,2	13,4	17,9
Indian or Asian	Female	0-4	15,5	4,4	11,1	19,9
Indian or Asian	Female	5- 9	14,7	2,6	12,1	17,3
Indian or Asian	Female	10-14	14,7	1,1	13,7	15,8
Indian or Asian	Female	15-19	15,4	1,6	13,9	17,0
Indian or Asian	Female	20-29	16,9	1,6	15,2	18,5
Indian or Asian	Female	30-44	18,6	2,5	16,1	21,1
Indian or Asian	Female	45-64	16,5	3,0	13,5	19,5
Indian or Asian	Female	65+	14,9	7,8	7,1	22,7
White	Male	0-4	21,5	3,1	18,3	24,6
White	Male	5-9	25,2	3,9	21,4	29,1
White	Male	10-14	25,1	5,5	19,6	30,7
White	Male	15-19	25,0	4,0	21,0	28,9
White	Male	20-29	25,6	1,8	23,8	27,4
White	Male	30-44	27,1	2,0	25,1	29,2
White	Male	45-64	21,9	2,6	19,2	24,5
White	Male	65+	18,7	1,7	16,9	20,4
White	Female	0-4	21,4	5,3	16,0	26,7
White	Female	5-9	25,2	7,2	17,9	32,4
White	Female	10-14	25,1	3,5	21,6	28,6
White	Female	15-19	24,9	3,2	21,7	28,0
White	Female	20-29	22,8	1,7	21,1	24,5
White	Female	30-44	26,8	3,0	23,8	29,9
White	Female	45-64	21,3	3,3	18,1	24,6
White	Female	65+	17,8	3,6	14,2	21,4

*subject to rounding error

2.3 The adjustment

The adjusted census population corresponds to the dual-system estimate of the true population. The actual adjustment procedure consisted of creating homogeneous adjustment classes with similar coverage rates within province – based on geography type, population group, sex, and age group – and calculating a common adjusted population, undercount rate, and adjustment factor, for each class separately. The national adjusted population was obtained by summing the adjusted classes. Only the population within the scope of the PES received adjustment factors. The totals for the balance of population (namely, people living in collective living quarters other than hostels, the homeless on the street, and those living in out-of-scope EAs) were not adjusted (see Sections 8.5 and 8.6).

It is the nature of statistical data to contain error. The adjusted population figures should always be analyzed with the full understanding that there is a certain degree of statistical uncertainty, that is, a range of possible values around them. They are subject to both sampling error (mainly random error) and non-sampling error (mainly biases). When comparing the census population figures with other sources of data, for example, demographic models and projections, the user must bear in mind statistical error, not only around the census figures, but around the model and projection estimates as well.

Table 2.7 shows the adjusted total population by geographic[1] and demographic classifications and the corresponding confidence intervals, which reflect the sampling error around the estimate. The confidence interval was obtained by adding the unadjusted balance of population to the lower and upper limits of the confidence interval for the adjusted population in the in-scope subuniverse.

Table 2.7
Adjusted total population – full universe
(in millions rounded to nearest thousand)

			95% Confidence interval limits	
Category	**Estimate**	**Absolute error (+ or -)**	**Lower**	**Upper**
All persons	44 820 000	392 000	44 428 000	45 212 000
Province				
Eastern Cape	6 437 000	150 000	6 286 000	6 587 000
Free State	2 707 000	41 000	2 665 000	2 748 000
Gauteng	8 837 000	317 000	8 520 000	9 154 000
KwaZulu-Natal	9 426 000	395 000	9 031 000	9 821 000
Limpopo	5 274 000	29 000	5 244 000	5 303 000
Mpumalanga	3 123 000	41 000	3 082 000	3 164 000
North West	3 669 000	61 000	3 608 000	3 731 000
Northern Cape	823 000	11 000	812 000	833 000
Western Cape	4 524 000	85 000	4 439 000	4 610 000
Population group				
Black African	35 416 000	493 000	34 923 000	35 909 000
Coloured	3 995 000	77 000	3 917 000	4 072 000
Indian or Asian	1 115 000	31 000	1 085 000	1 146 000
White	4 294 000	88 000	4 205 000	4 382 000.

[1] PES tables for urban/non-urban splits will be produced only after the revised definition is finalised by Stats SA.

Category	Estimate	Absolute error (+ or -)	95% Confidence interval limits	
			Lower	Upper
Sex (gender)				
Male	21 434 000	251 000	21 183 000	21 685 000
Female	23 386 000	277 000	23 109 000	23 663 000
Age group				
Under 5 years	4 450 000	59 000	4 391 000	4 509 000
5-9 years	4 854 000	75 000	4 779 000	4 928 000
10-14 years	5 062 000	74 000	4 988 000	5 136 000
15-19 years	4 982 000	61 000	4 920 000	5 043 000
20-29 years	8 229 000	96 000	8 133 000	8 325 000
30-44 years	9 032 000	106 000	8 926 000	9 138 000
45-64 years	5 996 000	65 000	5 931 000	6 061 000
65 or more years	2 215 000	24 000	2 192 000	2 239 000

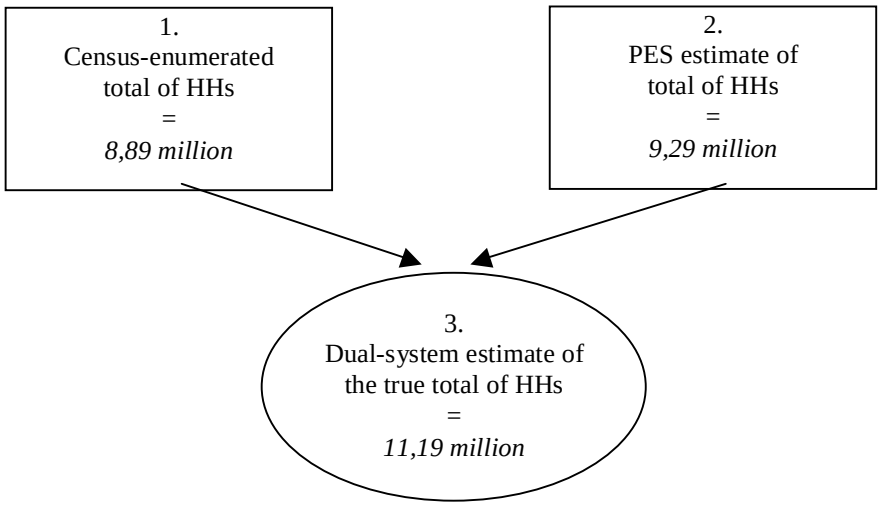
3 COVERAGE EVALUATION OF CENSUS 2001 – HOUSEHOLDS

The same dual-system estimation procedure described for persons in Section 2.1, and explained in detail in Section 8.3, was applied to households.

In Census 2001, a 'household' corresponds to the collection of persons in one questionnaire set (including continuation questionnaires). 'Questionnaire (parent questionnaire coupled with continuation questionnaires)' and 'household' thus refer to the same set of persons. Even though the basic definition for household is similar in both the census and PES, there are conceptual differences because the 'questionnaire' is not a fixed entity in the universe: the number of questionnaires completed for one housing unit can vary from interview to interview, especially in *de facto* enumerations which are based on presence rather than usual place of residence.

3.1 Estimation of true population

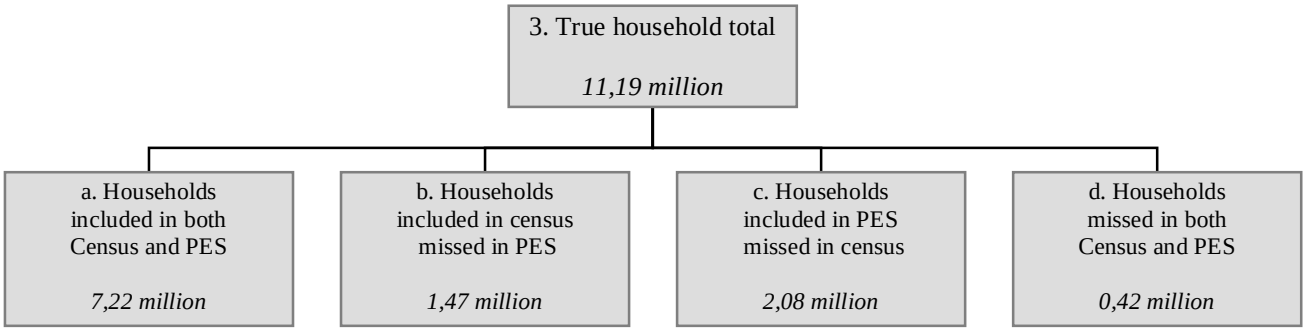
Figure 3.1
Estimates of total households from individual systems and from the dual system – in-scope subuniverse



In the subuniverse in scope for the PES, the separate census and PES enumerations produced 8,89 million and 9,29 million households, respectively. Using the dual-system estimation method, the true household total of South Africa in the in-scope subuniverse was estimated at 11,19 million.

Four components together make up the dual-system estimate of the true population.

Figure 3.2
Breakdown of dual-system estimate of household total – in-scope subuniverse



Note: Sums are subject to rounding error.

Component (a), the households included in both the census and the PES, was estimated at 7,22 million; component (b), the households included in the census but missed in the PES, was estimated at 1,47 million; component (c), the households included in the PES but missed in the census, was estimated at 2,08 million; and component (d), the households missed in both the census and the PES, was estimated at 0,42 million.

Table 3.1 shows that, of the 8,89 million households counted in the census for the in-scope subuniverse, 8,69 million are estimated to be correctly enumerated. Of these, the PES enumerated 7,22 million and missed 1,47 million. The census erroneous inclusions are estimated to be 0,20 million or 2,2% of the census total, approximately.

Table 3.1
Coverage distribution of Census households total – in-scope subuniverse
(in millions rounded to two decimals)

	Census enumeration
Total excluding erroneous inclusions	8,69
Included in PES	7,22
Omitted from PES	1,47
Erroneous inclusions	0,20
Total including erroneous inclusions	**8,89**

It is estimated that the census omitted 2,50 million households in total, 2,08 million of which were correctly enumerated in the PES, and another 0,42 million of which were missed in the PES as well as in the census (Table 3.2). This total omission does not take into account what it added incorrectly (the erroneous inclusions). When it is offset by the 0,20 million erroneous inclusions, the net undercount is 2,30 million. The net undercount relative to the 11,19 million in the true household total is thus approximately 20,55% (Table 3.5).

Table 3.2
Coverage distribution of true household total – in-scope subuniverse
(in millions rounded to two decimals)

		Census enumeration		
		Included	Omitted	**Total**
PES enumeration	Included	7,22	2,08	**9,29**
	Omitted	1,47	0,42	**1,90**
	Total excluding erroneous inclusions	**8,69**	**2,50**	**11,19**

Note: Sums are subject to rounding error.

While the PES estimated the household total in the in-scope subuniverse at 9,29 million, it omitted 1,47 million households that were correctly enumerated in the census, and another 0,42 million that were missed in both the census and the PES, for a total omission of 1,90 million (Table 3.2).

The true household total, estimated at 11,78 million, was calculated by adding the census-enumerated 0,59 million households in the balance of universe to the dual-system estimate of 11,19 million in the in-scope subuniverse (Table 3.3).

Table 3.3
Unadjusted and adjusted census household totals – full universe
(in millions rounded to two decimals)

	Households in housing units and hostels within in-scope EA types	**Households in other collective living quarters and other EA types**	**Total households**
Unadjusted	8,89	0,59	9,49
Adjusted	11,19	0,59	11,78

The overall empirical probabilities of inclusion and omission of a household in the census or in the PES are shown below in Table 3.4. A household in the in-scope universe had approximately a 77,7% chance of being enumerated in the census, an 83,0% chance of being enumerated in the PES, and a 63,6% chance of being enumerated in both. Conversely, the household had approximately a 13,2% chance of being included in the census but missed in the PES, an 18,6% chance of being included in the PES but missed in the census, and a 3,8% chance of being missed in both.

Table 3.4
Probabilities of inclusion and omission of a household – in-scope subuniverse

Probability of being included in the census	0,7766
Probability of being included in the PES	0,8303
Probability of being included in both the census and the PES	0,6448
Probability of being included in census, but missed in the PES	0,1318
Probability of being included in the PES, but missed in the census	0,1855
Probability of being missed in both the census and the PES	0,0379

3.2 Estimation of the net undercount rate

The net undercount (or overcount) is the difference between the estimated true household total and the census-enumerated household total. The rate is the net undercount expressed as a percentage of the estimated true total.

Net undercount rates, together with their absolute errors and confidence intervals, are shown in Table 3.5 for households by geographic[1] classification. The confidence interval is formed around the estimate by adding or subtracting the absolute error. See Section 2.2 for notes concerning confidence intervals and absolute errors, and their use in determining any differential undercounts.

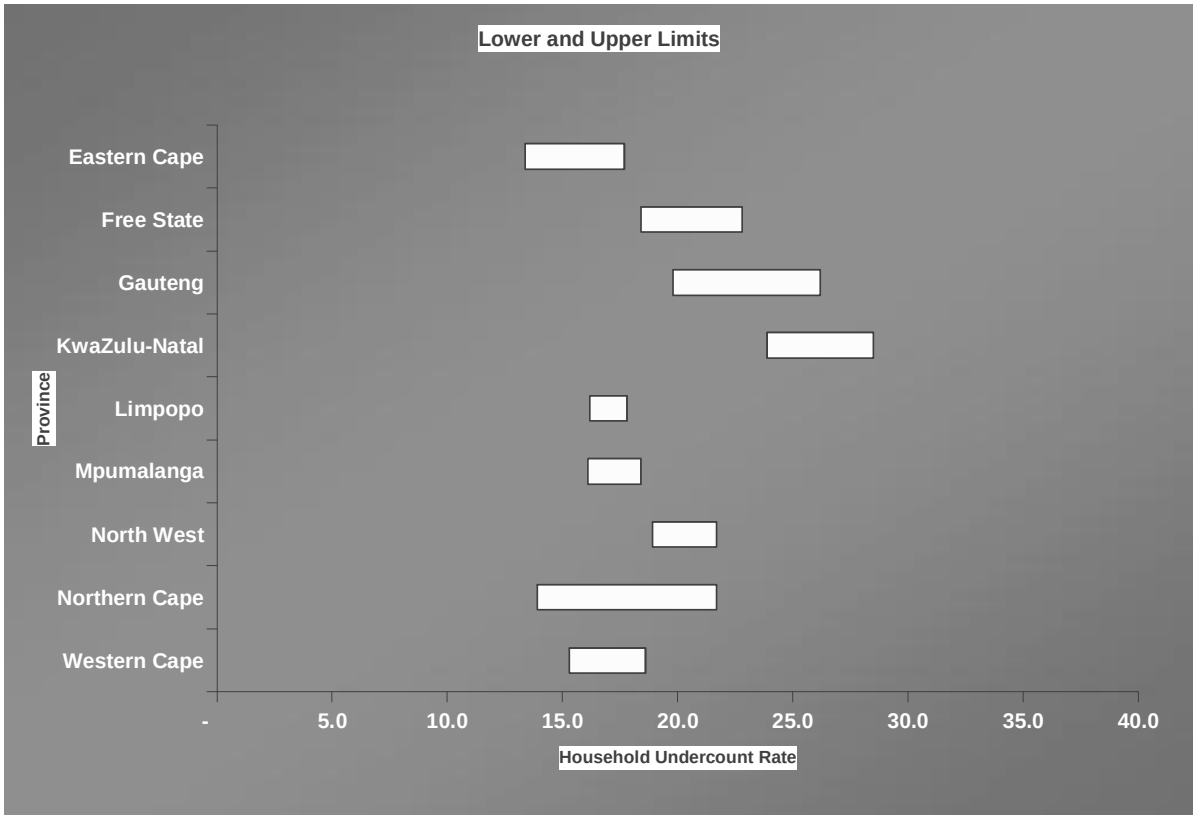
In Table 3.5, it can be observed that the net undercount rate for households at the national level was estimated at 20,5%, with possible values ranging from 19,5% to 21,5%.

[1] PES tables for urban/non-urban splits will be produced only after the revised definition is finalised by Stats SA.

Table 3.5
Net undercount rate for households by province – in-scope subuniverse

Category	Estimate	Absolute error (+ or -)	95% Confidence interval limits	
			Lower	Upper
All households	20,5	1,0	19,5	21,5
Province				
Eastern Cape	15,6	2,1	13,4	17,7
Free State	20,6	2,2	18,4	22,8
Gauteng	23,0	3,2	19,8	26,2
KwaZulu-Natal	26,2	2,3	23,9	28,5
Limpopo	17,0	0,8	16,2	17,8
Mpumalanga	17,2	1,2	16,1	18,4
North West	20,3	1,4	18,9	21,7
Northern Cape	17,8	3,9	13,9	21,7
Western Cape	16,9	1,6	15,3	18,6

Figure 3.3
Graphic representation of confidence intervals for household undercount rate – provinces



Among the provinces, the highest household undercount was observed in KwaZulu-Natal and Gauteng (26,2% and 23,0% respectively) (see Table 3.5 and Figure 3.3). The undercount in KwaZulu-Natal is significantly higher than in all provinces except Gauteng. There is no significant undercount difference between Gauteng and KwaZulu-Natal. Undercounts in the provinces of Eastern Cape, Limpopo, Mpumalanga, Northern Cape and Western Cape are not significantly different from one another. The lowest observed undercount, in Eastern Cape, is significantly lower than that in Free State, Gauteng and KwaZulu-Natal but not significantly lower than in the other provinces.

3.3 The adjustment

The adjusted census household total corresponds to the dual-system estimate of the true households. The adjustment procedure for households was similar to the adjustment procedure for persons. It consisted of creating homogeneous adjustment classes with similar coverage rates – based on geography type, province, household size, and population group of head of household – and calculating a common adjusted population, undercount rate and adjustment factor, for each class separately. The national adjusted household total was obtained by summing across the adjustment classes. Only the households in the in-scope subuniverse received adjustment factors. The balance of the households (i.e., in non-institutional collective living quarters other than hostels and in the out-of-scope EAs) were not adjusted (see Sections 8.5 and 8.6).

Table 3,6 shows the adjusted population by geographic[1] classification and the corresponding confidence intervals, which reflect the sampling error around the estimate. This estimate includes households in **housing units** only. It includes the housing units in the in-scope EAs and the out-of-scope EAs; however, it excludes hostels and other collective living quarters. The confidence interval for this estimate was obtained by adding the unadjusted households in out-of-scope housing units to the lower and upper limits of the confidence interval for the adjusted household total in in-scope housing units.

[1] PES tables for urban/non-urban splits will be produced only after the revised definition is finalised by Stats SA.

Table 3.6
Adjusted household total – housing-units universe
(all figures rounded to the nearest thousand)

			95% Confidence interval limits	
Category	**Estimate**	**Absolute Error (+ or -)**	**Lower**	**Upper**
All households	11 206 000	98 000	11 108 000	11 303 000
Province				
Eastern Cape	1 513 000	33 000	1 479 000	1 546 000
Free State	733 000	14 000	720 000	747 000
Gauteng	2 651 000	101 000	2 550 000	2 753 000
KwaZulu-Natal	2 086 000	68 000	2 018 000	2 155 000
Limpopo	1 180 000	10 000	1 170 000	1 190 000
Mpumalanga	733 000	9 000	724 000	743 000
North West	929 000	18 000	911 000	947 000
Northern Cape	207 000	7 000	200 000	214 000
Western Cape	1 173 000	22 000	1 151 000	1 196 000

4 CONTENT EVALUATION OF CENSUS 2001 – PERSONS ONLY

4.1 Nature of content analysis

Content error, also known as response error, is defined as the deviation of the obtained value from the true value for a given characteristic. Depending on whether essential or transient conditions are involved, response error can be divided into response bias (systematic error) and response variance (variable error).

The PES is regarded as a replication, an independent re-interview of a sample from the census for the purpose of estimating variable error, not bias. The PES content error analysis measures **consistency**, not which answers are right or wrong, i.e., it measures how **differently** answers are reported between the census and the PES.

The following characteristics were selected for content error analysis:

- Sex
- Age group
- Relationship to head of household
- Marital status
- Population group
- Home language
- Highest level of education

To ensure comparability between the PES and the census, the same wording, response categories and precodes, and also the same concept definitions, were maintained in the PES.

First, estimated totals from the census and the PES, as reported in the census and as reported in the PES, are compared for matched persons for the selected characteristics. The number of cases in agreement in the universe is observable along the diagonal.

Variability between the census and the PES is then measured by means of four different indicators: the net difference rate, the index of inconsistency (simple and aggregate), the gross difference rate, and the rate of agreement. These measures and their confidence intervals are presented for the selected characteristics.

- **Net Difference Rate (NDR).** The net difference rate is the difference between the number of cases in the census and the number of cases in the PES that fall under each response category, relative to the total number of matched persons in all response categories.

- **Index of Inconsistency.** The index of inconsistency is the relative number of cases for which the response varied between the census and the PES. It is the ratio of the simple response variance to the total variance of the characteristic, including its variability in the population. It is calculated for each response category.

- **Gross Difference Rate (also Off-Diagonal Proportion).** The gross difference rate (GDR) is calculated for the characteristic as a whole. It is the number of discrepancies between the census responses and the PES responses relative to the total number of persons matched. It is equivalent to the sum of all cells off the diagonal, for all categories, or the complement of the sum of the diagonal cells.

- **Rate of Agreement.** The rate of agreement is the complement of the gross difference rate. A low rate of agreement indicates a high degree of variability, and vice versa.

Figure 3.4
Standards for the interpretation of the different content error measures

Measure	Low	Moderate	High
Index of inconsistency	< 20	20–50	> 50
Aggregate index of inconsistency	< 20	20–50	> 50
Absolute value of NDR relative to mean or proportion (NDR/P)	<0,01	0,01–0,05	>0,05

Source: 'Evaluating censuses of Population and Housing', ISP-TR-5, US Census Bureau, 1985

Important note

The estimated person totals shown in the content analysis tables do not coincide with the final census totals for each characteristic because:

- they are based only on the sample of census records in the PES and are, therefore, subject to sampling variability;
- they include only matched cases, not the full sample;
- the data are unedited, while the data in the final census totals are edited;
- they include only the in-scope subuniverse (consisting of housing units and hostels within in-scope EA types) while the final census totals include the full universe; and
- they are unadjusted while the final census totals are adjusted for coverage error.

The sole purpose of these totals is to compare the census responses with the PES responses for consistency/variability analysis purposes. They are not intended for sociodemographic analysis purposes; final census results should be used for such purposes. The data quality in the final census results is, to a certain extent, greatly improved over what the content analysis indicates due to more accurate data capturing (by automated scanning with rigorous quality control systems) and to sophisticated editing procedures.

4.2 Content analysis for sex

Is (the person) male or female?

Table 4.2a
Sex as reported in the Census and as reported in the PES

	Sex (PES)			
Sex (Census)	Male	Female	Undetermined	Total PES
Male	10 7	733	30 245	11 555 798
Female	665 212	12 37	28 465	13 064 708
Undetermined	96 965	104	147	201 795
Total Census	11 554 629	13 20	58 857	24 822 301

Table 4.2b
Net difference rate, index of inconsistency, and gross difference rate for sex

				Net difference rate			Index of inconsistency		
	Total consistent cases	**Total in census**	**Total in PES**	**Rate**	**95% Confidence interval limits**		**Index**	**95% Confidence interval limits**	
Response category					Lower	Upper		Lower	Upper
Male	88 139	94 342	94 362	-0,01	-0,12	0,10	12,37	12,15	12,59
Female	100 151	105 799	106 980	-0,59	-0,70	-0,47	12,40	12,18	12,63
Undetermined	1	1 668	467	0,60	0,55	0,64	100,27	96,02	104,71
Total	188 291	201 809	201 809	-	-	-			
Aggregated index of inconsistency							13,30	13,08	13,53
Gross difference rate =	6,70%	(off-diagonal proportion)							
Rate of agreement =	93,30%								

The characteristic sex shows a **low** level of inconsistency or variability (index < 20%) and can be expected to be reported more or less reliably and consistently from survey to survey.

4.3 Content analysis for age group

What is (the person's) date of birth and age in completed years?

Table 4.3a
Age group as reported in the Census and as reported in the PES

	Age group (PES)							
Age group (Census)	0-4	5-14	15-19	20-29	30-44	45-64	65+	Total PES
0-4 years	2 106 936	170 496	16 384	25 022	21 962	12 116	6 242	2 359 158
5-14 years	160 731	5 372 865	191 974	61 238	24 720	11 265	7 738	5 830 531
15-19 years	12 865	179 276	2 493 134	131 884	15 733	9 377	3 301	2 845 570
20-29 years	26 611	61 254	123 596	3 832 036	156 056	23 814	7 686	4 231 053
30-44 years	20 889	20 640	17 588	148 725	4 434 805	199 582	16 829	4 859 058
45-64 years	14 647	13 596	9 049	25 223	197 576	3 088 455	83 730	3 432 276
65+ years	6 631	10 806	4 327	10 559	16 214	84 207	1 131 912	1 264 656
Total census	2 349 310	5 828 933	2 856 052	4 234 687	4 867 066	3 428 816	1 257 438	24 822 302

Table 4.3b
Net difference rate, index of inconsistency, and gross difference rate for age group

				Net difference rate			Index of inconsistency		
	Total consistent cases	**Total in census**	**Total in PES**	**Rate**	**95% Confidence interval limits**		**Index**	**95% Confidence interval limits**	
Response category					Lower	Upper		Lower	Upper
0-4 years	17 038	19 068	18 974	0,05	-0,02	0,11	11,51	11,15	11,88
5-14 years	42 589	46 239	46 176	0,03	-0,05	0,12	10,16	9,92	10,40
15-19 years	19 955	22 741	22 853	-0,06	-0,13	0,02	14,05	13,69	14,43
20-29 years	31 518	34 733	34 763	-0,01	-0,09	0,06	11,23	10,95	11,51
30-44 years	37 079	40 565	40 629	-0,03	-0,11	0,05	10,85	10,59	11,11
45-64 years	25 690	28 455	28 467	-0,01	-0,08	0,07	11,33	11,03	11,64
65+ years	8 960	10 008	9 947	0,03	-0,01	0,07	10,73	10,26	11,21
Total	182 829	201 809	201 809	-	-	-			
Aggregated index of inconsistency							11,28	11,12	11,45
Gross difference rate =	9,40% (off-diagonal proportion)								
Rate of agreement =	90,60%								

In both the PES and the census, age was derived from the date of birth; in other words, age derived from the date of birth was preferred over reported age when the two were inconsistent. The characteristic age (as derived) shows a **low** level of inconsistency or variability (index < 20%) and can be expected to be reported more or less reliably and consistently from survey to survey.

4.4 Content analysis for relationship to head of household

What is (the person's) relationship to the head or acting head of the household?

Table 4.4a
Relationship to head of household as reported in the Census and as reported in the PES

	Relationship to head of household (PES)							
Relationship to head of household (Census)	Head/ acting head	Husband/ wife/ partner	Son/ daughter	Adopted child	Stepchild	Brother/ sister	Parent	Parent-in-law
Head/acting head	5 415 601	407 290	138 435	1 602	2 403	73 051	58 051	8 673
Husband/wife/partner	477 428	2 336 789	65 661	828	1 325	14 459	15 468	2 859
Son/daughter	123 934	64 940	8 269 686	10 868	57 533	188 575	87 090	7 376
Adopted child	1 671	604	24 150	6 880	2 278	1 259	-	-
Stepchild	1 495	762	54 978	1 458	10 679	4 234	1 240	-
Brother/sister	73 058	15 879	227 924	679	2 966	409 705	3 756	2 653
Parent	93 105	26 287	49 931	33	558	3 014	71 754	13 369
Parent-in-law	10 706	3 401	8 500	459	195	2 366	12 140	21 552
Grand/great-grandchild	20 129	5 757	364 354	7 149	10 941	29 026	11 426	4 450
Son/daughter-in-law	9 179	9 009	51 125	582	2 650	9 281	1 933	4 391
Brother/sister-in-law	9 356	6 407	27 505	886	666	28 014	-	1 314
Other relative	33 411	16 300	136 694	4 433	3 213	48 148	5 864	6 266
Non-related person	26 069	9 430	20 712	2 022	1 590	8 297	1 073	834
Undetermined	47 443	10 853	53 403	113	600	4 414	869	192
Total census	6 342 585	2 913 708	9 493 058	37 992	97 597	823 843	270 664	73 929

Continued…

	Relationship to head of household (PES)						
Relationship to head of household (Census)	Grand/ great-grandchild	Son/ daughter-in-law	Brother/ sister-in-law	Other relative	Non-related person	Undetermined	Total PES
Head/acting head	24 646	12 515	8 511	43 597	18 766	4 424	6 217 565
Husband/wife/partner	10 190	16 672	5 575	20 676	12 188	2 582	2 982 700
Son/daughter	393 221	37 812	16 949	200 004	19 051	24 325	9 501 364
Adopted child	6 951	-	275	8 982	3 140	113	56 303
Stepchild	5 089	1 374	392	12 144	930	270	95 045
Brother/sister	35 607	6 850	26 439	77 372	10 295	1 292	894 475
Parent	13 967	2 035	1 259	10 448	1 458	223	287 441
Parent-in-law	5 577	8 019	820	7 168	1 360	-	82 263
Grand/great-grandchild	2 599 533	19 401	5 466	151 353	6 527	15 339	3 250 851
Son/daughter-in-law	23 372	74 369	8 038	38 126	4 517	402	236 974
Brother/sister-in-law	10 204	13 342	29 511	40 677	4 862	144	172 888
Other relative	104 379	15 144	17 068	284 402	32 079	3 557	710 958
Non-related person	7 486	2 064	2 652	34 774	65 168	745	182 916
Undetermined	20 133	623	1 845	7 610	2 330	129	150 557
Total census	3 260 355	210 220	124 800	937 333	182 671	53 545	24 822 300

Table 4.4b
Net difference rate, index of inconsistency, and gross difference rate for relationship to head of household

				Net difference rate			Index of inconsistency		
	Total consistent cases	Total in Census	Total in PES	Rate	95% confidence interval limits		Index	95% confidence interval limits	
Response category					Lower	Upper		Lower	Upper
Head/acting head	44 485	50 871	51 927	-0,52	-0,64	-0,41	18,05	17,74	18,36
Husband/wife/partner	19 973	25 214	24 630	0,29	0,19	0,39	22,66	22,20	23,12
Son/daughter	66 938	76 741	76 663	0,04	-0,10	0,18	20,53	20,24	20,83
Adopted child	73	496	341	0,08	0,05	0,10	82,72	76,66	89,26
Stepchild	99	837	813	0,01	-0,03	0,05	88,36	83,84	93,12
Brother/sister	3 238	6 967	6 511	0,23	0,14	0,31	53,74	52,47	55,04
Parent	565	2 218	2 296	-0,04	-0,10	0,02	75,81	73,25	78,47
Parent-in-law	181	645	598	0,02	-0,01	0,05	71,10	66,46	76,05
Grand/great-grandchild	20 616	25 909	25 713	0,10	0,00	0,20	23,08	22,63	23,54
Son/daughter-in-law	556	1 764	1 582	0,09	0,04	0,14	67,32	64,53	70,23
Brother/sister-in-law	262	1 402	1 028	0,19	0,14	0,23	78,90	75,37	82,60
Other relative	2 407	5 880	7 622	-0,86	-0,96	-0,77	66,53	65,12	67,98
Non-related person	606	1 616	1 664	-0,02	-0,07	0,02	63,57	60,83	66,42
Undetermined	1	1 249	421	0,41	0,37	0,45	100,19	95,41	105,22
Total	**160 000**	**201 809**	**201 809**	-	-	-			
Aggregated index of inconsistency							27,38	27,15	27,62
Gross difference rate =	20,72%	(off-diagonal proportion)							
Rate of Agreement =	79,28%								
NOTE: this variable may not represent the same person in both the Census and the PES.									

The characteristic ‘relationship to head of household’ shows a **moderate** level of inconsistency or variability (20% < index < 50%). It may not be reported consistently from survey to survey. In the case of Census 2001 and PES 2001, the inconsistencies are most likely due to the fact that the person referred to as head was not necessarily the same in both cases. Given the *de facto* enumeration rule (based on presence), the person could have been the ‘acting head’ in the absence of the normal head of household. It may not be possible to ensure more consistent responses for this variable in future surveys unless a *de jure* rule (usual residence) is used. With a *de jure* rule, the head of household generally remains the same, even when temporarily absent from the household.

4.5 Content analysis for marital status

What is (the person's) PRESENT marital status?

Table 4.5a
Marital status as reported in the Census and as reported in the PES

	Marital status (PES)				
Marital status (Census)	Married civil/religious	Married traditional/ customary	Polygamous marriage	Living together like married partners	Never married
Married civil/religious	3 260 677	398 532	4 003	61 782	140 681
Married traditional/customary	339 653	1 061 773	6 480	133 961	124 068
Polygamous marriage	8 279	11 187	2 890	1 150	5 215
Living together like married partners	91 507	200 368	1 750	681 513	201 656
Never married	229 024	143 856	11 207	218 516	14 377 199
Widower/widow	65 176	71 580	578	13 125	105 168
Separated	19 110	13 539	227	6 160	62 323
Divorced	19 313	8 129	-	9 247	61 493
Undetermined	34 084	14 548	664	7 150	939 057
Total census	4 066 823	1 923 512	27 799	1 132 604	16 016 860

Continued…

	Marital status (PES)				
Marital status (Census)	Widower/ widow	Separated	Divorced	Undeter-mined	Total PES
Married civil/religious	82 677	24 168	16 087	3 433	3 992 040
Married traditional/customary	76 490	18 572	7 256	2 807	1 771 060
Polygamous marriage	1 165	343	353	-	30 582
Living together like married partners	9 642	3 773	8 099	2 694	1 201 002
Never married	115 828	60 180	66 305	49 614	15 271 729
Widower/widow	762 995	16 291	18 721	1 880	1 055 514
Separated	15 586	56 156	24 305	270	197 676
Divorced	20 217	27 898	144 664	253	291 214
Undetermined	5 001	1 764	2 078	7 142	1 011 488
Total census	1 089 601	209 145	287 868	68 093	24 822 305

Table 4.5b
Net difference rate, index of inconsistency, and gross difference rate for marital status

				Net difference rate			Index of inconsistency		
	Total consistent cases	**Total in Census**	**Total in PES**	**Rate**	**95% confidence interval limits**		**Index**	**95% confidence interval limits**	
Response category					Lower	Upper		Lower	Upper
Married civil/religious	29 101	35 018	35 566	-0,27	-0,38	-0,16	21,26	20,88	21,65
Married traditional/ customary	7 374	12 652	13 908	-0,62	-0,73	-0,51	47,60	46,73	48,48
Polygamous marriage	20	221	211	0,00	-0,01	0,02	90,84	82,11	100,49
Living together like married partners	6 243	10 459	9 999	0,23	0,14	0,32	41,05	40,14	41,98
Never married	115 714	122 985	128 987	-2,97	-3,12	-2,83	21,68	21,40	21,97
Widower/widow	6 095	8 347	8 593	-0,12	-0,19	-0,05	29,27	28,43	30,13
Separated	439	1 522	1 634	-0,06	-0,10	-0,01	72,75	69,76	75,86
Divorced	1 250	2 406	2 387	0,01	-0,04	0,06	48,42	46,44	50,48
Undetermined	53	8 199	524	3,80	3,71	3,90	99,27	97,15	101,43
Total	**166 289**	**201 809**	**201 809**	-	-	-			
Aggregated index of inconsistency							30,82	30,53	31,12
Gross difference rate =	17,60%	(off-diagonal proportion)							
Rate of Agreement =	82,40%								

The characteristic 'marital status' shows a **moderate** level of inconsistency or variability (20% < index < 50%). It may not be reported consistently from survey to survey. At the level of each response category, most categories show an even **greater** degree of inconsistency. The most stable response seems to be 'married civil/religious' followed by 'never married'. Even for those, the degree of variability is not low. To ensure more reliable responses in future censuses and surveys, different choices of wording and instructions will have to be tested. More probing may also be needed.

4.6 Content analysis for population group

How would (the person) describe him/herself in terms of population group?

Table 4.6a
Population group as reported in the Census and as reported in the PES

	Population group (PES)						
Population group (Census)	Black African	Coloured	Indian or Asian	White	Other	Undeter-mined	Total PES
Black African	19 935 533	111 173	3 775	36 900	19 898	52 095	20 159 374
Coloured	79 378	2 217 506	9 894	8 505	7 304	6 106	2 328 693
Indian or Asian	9 594	13 792	503 180	3 014	3 483	1 119	534 182
White	42 638	6 320	4 188	1 505 751	3 612	5 294	1 567 803
Other	25 857	17 068	4 359	3 213	4 132	135	54 764
Undetermined	130 253	24 784	3 184	18 000	134	1 132	177 487
Total census	20 223 253	2 390 643	528 580	1 575 383	38 563	65 881	24 822 303

Table 4.6b
Net difference rate, index of inconsistency, and gross difference rate for population group

				Net difference rate			Index of inconsistency		
	Total consistent cases	**Total in Census**	**Total in PES**	**Rate**	**95% Confidence interval limits**		**Index**	**95% Confidence interval limits**	
Response category					Lower	Upper		Lower	Upper
Black African	153 317	155 463	155 947	-0,24	-0,31	-0,17	6,71	6,52	6,91
Coloured	24 556	25 811	26 359	-0,27	-0,33	-0,22	6,73	6,49	6,98
Indian or Asian	4 230	4 498	4 453	0,02	0,00	0,04	5,61	5,13	6,14
White	13 500	14 054	14 164	-0,05	-0,09	-0,02	4,64	4,38	4,91
Other	41	490	383	0,05	0,03	0,08	90,80	84,57	97,49
Undetermined	9	1 493	503	0,49	0,45	0,53	99,47	95,10	104,04
Total	195 653	201 809	201 809	-	-	-			
Aggregated index of inconsistency							7,97	7,77	8,18
Gross difference rate =	3,05% (off-diagonal proportion)								
Rate of agreement =	96,95%								

The characteristic 'population group' exhibits the **lowest** overall degree of inconsistency and variability among the characteristics measured. It seems to be quite robust and reliable from one measurement to another. At the individual response category level, the two categories 'other' and 'undetermined' do show great inconsistency from census to PES. However, they only occur in a few cases and in the final census results they are edited out.

4.7 Content analysis for home language

Which language does (the person) speak most often in this household?

Table 4.7a
Home language as reported in the Census and as reported in the PES

	Language (PES)						
Language (Census)	Afrikaans	English	IsiNdebele	IsiXhosa	IsiZulu	Sepedi	Sesotho
Afrikaans	2 731 966	166 482	2 265	25 110	7 835	2 215	8 214
English	169 623	1 417 918	3 181	10 890	18 399	1 684	1 404
IsiNdebele	113	2 337	268 381	2 482	51 320	43 944	6 474
IsiXhosa	25 919	7 885	3 379	4 344 657	61 179	3 272	69 878
IsiZulu	7 518	17 727	38 406	68 405	5 107 754	21 414	71 160
Sepedi	2 183	1 261	16 849	3 032	25 823	1 869 231	156 121
Sesotho	4 678	2 680	3 430	41 755	76 941	58 304	2 106 575
Setswana	11 409	2 227	3 573	13 282	23 993	40 018	87 210
SiSwati	525	157	1 993	1 561	58 936	8 426	10 837
Tshivenda	1 541	98	144	1 571	5 280	11 568	4 017
Xitsonga	2 899	313	457	14 667	22 709	21 634	8 006
Other	1 781	18 961	475	1 128	5 046	801	1 144
Undetermined	22 952	13 017	4 013	30 989	44 482	10 634	15 712
Total Census	**2 983 107**	**1 651 063**	**346 546**	**4 559 529**	**5 509 697**	**2 093 145**	**2 546 752**

continued…

Continued…

	Language (PES)						
Language (Census)	Setswana	SiSwati	Tshivenda	Xitsonga	Other	Undetermined	**Total PES**
Afrikaans	10 393	295	1 357	1 793	1 652	7 867	**2 967 444**
English	3 631	-	253	1 613	24 667	4 526	**1 657 789**
IsiNdebele	6 597	1 458	512	393	675	2 415	**387 101**
IsiXhosa	20 918	2 142	2 729	10 821	3 269	12 967	**4 569 015**
IsiZulu	25 578	45 375	6 435	18 444	2 691	17 576	**5 448 483**
Sepedi	56 372	9 933	2 681	16 578	945	8 449	**2 169 458**
Sesotho	69 493	3 304	5 127	6 592	2 346	3 449	**2 384 674**
Setswana	2 038 033	6 014	2 422	11 315	941	3 902	**2 244 339**
SiSwati	12 024	639 760	1 075	16 370	688	2 923	**755 275**
Tshivenda	4 194	230	805 638	10 214	552	856	**845 903**
Xitsonga	22 980	8 176	10 763	1 025 488	9 670	3 243	**1 151 005**
Other	1 191	249	1 851	10 366	27 003	144	**70 140**
Undetermined	12 696	7 764	2 581	5 150	666	1 014	**171 670**
Total Census	**2 284 100**	**724 700**	**843 424**	**1 135 137**	**75 765**	**69 331**	**24 822 296**

Table 4.7b
Net difference rate, index of inconsistency, and gross difference rate for home language

				Net difference rate			Index of inconsistency		
	Total consistent cases	**Total in Census**	**Total in PES**	**Rate**	**95% Confidence interval limits**		**Index**	**95% Confidence interval limits**	
Response category					Lower	Upper		Lower	Upper
Afrikaans	31 314	33 455	33 684	-0,11	-0,18	-0,05	8,06	7,82	8,30
English	11 561	13 504	13 498	0,00	-0,06	0,06	15,40	14,91	15,90
IsiNdebele	2 661	3 614	3 371	0,12	0,08	0,16	24,23	23,07	25,44
IsiXhosa	29 093	30 925	30 883	0,02	-0,04	0,08	6,92	6,69	7,15
IsiZulu	40 911	43 668	44 174	-0,25	-0,33	-0,17	8,76	8,54	8,99
Sepedi	13 721	16 172	15 465	0,35	0,29	0,41	14,39	13,95	14,84
Sesotho	16 277	18 507	19 776	-0,63	-0,70	-0,55	16,53	16,10	16,97
Setswana	18 007	19 860	20 167	-0,15	-0,21	-0,09	11,13	10,78	11,49
SiSwati	6 030	7 035	6 744	0,14	0,10	0,19	12,92	12,31	13,55
Tshivenda	5 020	5 343	5 318	0,01	-0,01	0,04	5,98	5,52	6,48
Xitsonga	6 688	7 677	7 547	0,06	0,02	0,11	12,61	12,04	13,22
Other	232	623	656	-0,02	-0,04	0,01	63,92	59,60	68,56
Undetermined	8	1 426	526	0,45	0,40	0,49	99,56	95,14	104,19
Total	**181 523**	**201 809**	**201 809**	-	-	-			
Aggregated index of inconsistency							11,57	11,42	11,73
Gross difference rate =	10,05%	(off-diagonal proportion)							
Rate of agreement =	89,95%								

The characteristic ‘home language’ shows a **low** level of inconsistency or variability (index < 20%). It can be expected to be reported more or less reliably and consistently from survey to survey. The few inconsistencies could be due to people confusing the language spoken most often in the household with their mother tongue, or to more than one language being spoken in the household. At the individual response category level, the languages reported with the highest consistency were: Tshivenda, isiXhosa, Afrikaans and isiZulu.

4.8 Content analysis for highest level of education

What is the highest level of education that (the person) has completed?

Table 4.8a
Highest level of education as reported in the Census and as reported in the PES

	Highest level of education (PES)							
Highest level of education (Census)	No schooling	Grade 0	Grade 1/ Sub A	Grade 2/ Sub B	Grade 3/ Std 1	Grade 4/ Std 2	Grade 5/ Std 3	Grade 6/ Std 4
No schooling	272 329	88 610	35 191	14 579	8 873	7 016	7 427	6 080
Grade 0	210 944	287 687	125 918	47 018	17 819	10 743	9 180	8 401
Grade 1/Sub A	39 391	258 873	391 102	152 594	58 471	29 158	20 697	13 618
Grade 2/Sub B	22 601	52 889	303 659	415 251	157 751	60 049	36 211	23 897
Grade 3/Std 1	7 859	19 870	60 442	297 190	415 032	165 066	72 496	34 217
Grade 4/Std 2	5 165	10 636	32 797	75 547	298 688	457 152	213 682	90 950
Grade 5/Std 3	6 424	8 097	18 371	42 851	72 471	348 470	617 473	252 103
Grade 6/Std 4	4 244	6 772	13 762	23 623	40 620	85 183	420 476	732 332
Grade 7/Std 5	5 616	4 347	8 890	11 735	15 503	34 755	80 367	342 625
Grade8/Std 6/Form 1	5 204	4 062	7 110	10 599	14 400	19 963	54 597	105 978
Grade9/Std 7/Form 2	3 874	1 878	3 449	5 027	7 389	9 160	20 321	33 320
Gr 10/Std 8/ Form 3/NTCI	3 331	4 255	5 367	8 826	10 252	15 974	25 543	32 394
Gr 11/Std 9/Form/NTCII	135	139	140	129	105	257	553	699
Gr 12/Std10/Form/Matric./ NTCIII	144	-	-	52	-	-	347	151
Certificate w/ less than Gr 12	631	988	129	957	1 189	1 448	1 252	2 302
Diploma w/ less than Gr 12	332	592	334	144	876	1 366	1 419	1 750
Certificate with Grade 12	135	343	-	-	144	529	576	837
Diploma with Grade 12	-	-	-	-	-	118	118	609
Bachelors Degree			125	135	-	262	-	247
Bachelors Deg and Diploma		195	-	-	144	-	273	273
Honours degree	87	-	144	179	276	-	705	243
Higher Degree (Mast/Doct.)	1 546	2 046	5 544	6 625	8 943	9 904	11 444	11 385
Undetermined	180 348	116 652	144 465	157 205	148 123	158 740	180 329	175 158
Total Census	770 340	868 931	1 156 939	1 270 266	1 277 069	1 415 313	1 775 486	1 869 569

Continued…

	Highest level of education (PES)							
Highest level of education (Census)	Grade 7/ Std 5	Grade8/ Std 6/ Form 1	Grade9/ Std 7/ Form 2	Grade 10/ Std 8/ Form 3/ NTCI	Grade 11/ Std 9/ Form 4/ NTCII	Grade 12/ Std10/ Form/ Matric./ NTCIII	Certificate with less than Grade 12	Diploma with less than Grade 12
No schooling	4 745	5 424	2 761	5 079	-	-	-	231
Grade 0	4 945	6 593	2 976	4 852	-	-	144	171
Grade 1/Sub A	8 671	6 956	2 823	6 582	-	-	-	817
Grade 2/Sub B	9 311	12 552	6 378	8 503	291	280	440	612
Grade 3/Std 1	15 707	14 800	6 581	10 335	357	262	773	414
Grade 4/Std 2	34 396	25 688	10 848	14 142	306	-	481	2 016
Grade 5/Std 3	88 081	45 092	17 345	19 191	713	-	1 245	1 392
Grade 6/Std 4	207 022	122 040	30 285	38 243	1 466	706	1 475	4 841
Grade 7/Std 5	518 004	233 030	60 875	41 220	663	860	2 342	2 622
Grade8/Std 6/Form 1	360 707	752 578	195 927	109 253	3 895	3 107	5 223	8 336
Grade9/Std 7/Form 2	64 814	295 161	537 454	198 739	3 264	1 161	7 168	7 994
Gr 10/Std 8/ Form 3/ NTCI	44 797	117 407	309 327	1 851 564	12 802	8 710	87 499	141 227
Gr 11/Std 9/Form 4/NTCII	1 650	3 420	3 112	15 264	1 101	170	2 461	2 922
Gr 12/Std10/Form5/Matric./ NTCIII	248	2 163	1 000	11 342	617	1 079	1 808	10 968
Certificate w/ less than Gr 12	1 638	6 092	9 694	102 704	2 375	2 467	22 977	34 929
Diploma w/ less than Gr 12	2 253	7 777	5 573	110 486	3 136	11 405	24 972	227 105
Certificate with Grade 12	243	674	1 181	23 235	279	316	2 578	17 236
Diploma with Grade 12	140	542	343	7 315	118	649	717	14 305
Bachelors Degree	516	270	144	4 281	195	135	325	1 791
Bachelors Deg and Diploma	262	558	-	3 034	-	257	279	2 052
Honours degree	659	1 064	974	4 798	379	432	899	3 585
Higher Degree (Mast/ Doct.)	6 602	9 628	4 918	7 640	294	279	113	888
Undetermined	122 613	124 774	102 419	174 992	2 553	1 783	11 500	27 772
Total Census	1 498 024	1 794 283	1 313 085	2 772 794	34 804	34 058	175 419	514 226

continued …

Continued…

	Highest level of education (PES)							
Highest level of education (Census)	Certificate with Grade 12	Diploma with Grade 12	Bachelors Degree	Bachelors Degree and Diploma	Honours Degree	Higher Degree (Masters Doctorate)	Undeter-mined	Total PES
No schooling	14	134	-	-	853	1 185	347 757	808 288
Grade 0	118	-	-	-	205	3 472	81 144	822 330
Grade 1/Sub A	171	280	378	144	1 056	5 225	72 838	1 069 845
Grade 2/Sub B	280	-	-	-	1 661	11 576	66 436	1 190 628
Grade 3/Std 1	144	-	-	-	1 122	10 502	50 924	1 184 093
Grade 4/Std 2	279	217	98	257	325	11 704	42 918	1 328 292
Grade 5/Std 3	757	579	177	171	940	17 583	53 370	1 612 896
Grade 6/Std 4	460	507	98	-	1 163	22 907	42 189	1 800 414
Grade 7/Std 5	923	663	215	98	1 956	9 591	33 563	1 410 463
Grade8/Std 6/Form 1	2 070	1 325	275	373	2 614	14 051	26 734	1 708 381
Grade9/Std 7/Form 2	1 888	253	233	33	2 454	7 244	18 380	1 230 658
Gr 10/Std 8/Form 3/NTCI	34 702	11 395	5 358	4 682	16 114	17 696	34 939	2 804 161
Gr 11/Std 9/Form 4/NTCII	470	127	-	135	277	348	968	34 582
Gr 12/Std10/Form/Matric./ NTCIII	1 150	315	288	-	476	492	262	32 902
Certificate w/ less than Gr 12	3 446	749	387	233	1 301	610	2 502	201 000
Diploma w/ less than Gr 12	24 711	14 283	3 649	2 153	6 833	3 207	2 374	456 730
Certificate with Grade 12	68 964	16 881	6 207	3 793	3 146	1 032	1 750	150 079
Diploma with Grade 12	22 868	14 838	4 659	2 373	1 854	597	313	72 476
Bachelors Degree	10 159	3 879	23 236	4 736	782	628	255	52 101
Bachelors Deg and Diploma	8 320	2 234	3 490	33 416	1 626	397	-	56 810
Honours Degree	2 162	596	411	748	2 120	558	2 728	23 747
Higher Degree (Mast/ Doct.)	515	171	-	144	1 577	12 377	19 208	121 791
Undetermined	7 999	2 185	1 679	1 885	8 854	48 052	4 749 405	6 649 485
Total census	192 570	71 611	50 838	55 374	59 309	201 034	5 650 957	24 822 299

Table 4.8b
Net difference rate, index of inconsistency, and gross difference rate for highest level of education

				Net difference rate			Index of inconsistency		
	Total consistent cases	**Total in Census**	**Total in PES**	**Rate**	**95% Confidence interval limits**		**Index**	**95% Confidence interval limits**	
Response category					Lower	Upper		Lower	Upper
No schooling	18 291	27 740	21 613	3,04	2,92	3,15	29,42	28,90	29,94
Grade 0	21 971	25 008	23 625	0,69	0,62	0,75	10,97	10,65	11,29
Grade 1/Sub A	4 706	6 331	6 068	0,13	0,08	0,18	24,85	23,96	25,78
Grade 2/Sub B	4 963	6 465	6 801	-0,17	-0,22	-0,11	26,03	25,15	26,95
Grade 3/Std 1	6 525	8 539	9 167	-0,31	-0,38	-0,24	27,50	26,71	28,32
Grade 4/Std 2	7 002	9 611	10 193	-0,29	-0,36	-0,21	30,80	30,00	31,62
Grade 5/Std 3	7 090	9 612	10 319	-0,35	-0,43	-0,28	30,35	29,56	31,16
Grade 6/Std 4	7 961	10 930	11 481	-0,27	-0,35	-0,19	30,66	29,90	31,43
Grade 7/Std 5	10 365	13 262	14 538	-0,63	-0,72	-0,55	27,31	26,67	27,97
Grade8/Std 6/Form 1	11 004	14 816	15 323	-0,25	-0,34	-0,16	29,15	28,52	29,81
Grade9/Std 7/Form 2	8 958	11 555	12 229	-0,33	-0,41	-0,26	26,22	25,54	26,91
Grade 10/Std 8/Form 3/NTCI	10 623	14 184	14 785	-0,30	-0,38	-0,21	28,72	28,07	29,38
Grade 11/Std 9/Form 4/NTCII	8 412	9 950	10 544	-0,29	-0,35	-0,23	18,86	18,25	19,50
Grade 12/Std10/Form/ Matric./ NTCIII	17 467	23 622	23 321	0,15	0,04	0,26	28,95	28,43	29,48
Certificate w/ less than Grade 12	13	299	292	0,00	-0,02	0,03	95,74	88,02	104,14
Diploma with less than Grade 12	36	275	306	-0,02	-0,04	0,01	87,73	80,29	95,86
Certificate with Grade 12	437	1 702	1 511	0,09	0,05	0,14	73,38	70,41	76,48
Diploma with Grade 12	2 412	3 957	4 412	-0,23	-0,28	-0,17	43,25	41,82	44,73
Bachelors Degree	979	1 253	1 608	-0,18	-0,21	-0,15	31,78	29,74	33,97
Bachelors Degree and Diploma	297	600	598	0,00	-0,02	0,03	50,57	46,62	54,85
Honours degree	270	443	431	0,01	-0,01	0,02	38,30	34,33	42,72
Higher Degree (Masters/ Doct.)	310	457	455	0,00	-0,02	0,02	32,09	28,55	36,07
Other	18	195	524	-0,16	-0,19	-0,14	95,13	88,12	102,69
Don't know	102	1 000	1 665	-0,33	-0,38	-0,28	--	--	--
Undetermined	0	3	0	0,00	0,00	0,00	100,00	51,20	282,14
Total	**150 212**	**201 809**	**201 809**	-	-	-			
Aggregated index of inconsistency							25,57	25,37	25,76
Gross difference rate =	25,57%	(off-diagonal-band proportion)							
Rate of agreement =	74,43%								

In the case of 'highest level of education,' compatible categories were grouped together (forward grouping for progression of educational level; only Certificate with less than Grade 12 has been grouped backward) for the total of consistent cases. By accepting values in a range around the diagonal as consistent, a diagonal band rather than a diagonal line was formed. The reason for combining categories is that people may have added 1 more year of schooling, since the PES was closer to the end of the school year than the census was.

With compatible categories of responses combined, the characteristic 'highest educational level' shows a **moderate** level of inconsistency or variability (20% < index < 50%). At the individual response category level, only Grade 0 and Grade 11 show a low level of inconsistency. However, when categories are not combined (table not shown), the overall degree of inconsistency is **high** (index > 50%). To ensure more reliable responses in future censuses and surveys, different choices of wording and instructions will have to be tested. More probing will be required as well.

4.9 Summary of content error analysis

The characteristics 'population group', 'age', 'home language', and 'sex' show a low level of variability between the census and the PES. They can be expected to be measured reliably from survey to survey. The variables 'relationship to head of household', 'marital status' and 'highest level of education (combining categories)' show a moderate level of variability, which might be indicative of a need for clearer concept definitions and wording, and more probing. The variable 'highest level of education' shows, in addition, sensitivity to the reference period, as evidenced by the increase in the inconsistency level when categories are not combined.

Table 4.9
Characteristics ranked from lowest to highest inconsistency

Characteristic	Aggregated index of inconsistency	Interpretation
Population group	7,97 %	low
Age	11,28 %	low
Home language	11,57 %	low
Sex	13,30 %	low
Highest level of education (combining categories)	25,57 %	moderate
Relationship to head of household	27,38 %	moderate
Marital status	30,82 %	moderate
Highest level of education (categories not combined)	55,47 %	high

Part II

METHODOLOGY AND PROCEDURAL HISTORY

5. ANALYTICAL OBJECTIVES

5.1 Domains of estimation

The target universe for the PES was discussed in Section 1.2. Specifically, PES estimates apply only to the housing-unit and hostel subuniverse in selected EA types; other collective living quarters are excluded from the scope of the PES. The domains of estimation for which reliable estimates can be expected with the given sample size (Section 6.4) are:

- National
- Urban/non-urban at national level
- Province

Within these domains, subpopulations are defined during estimation. Estimates are calculated separately for sex, age group, and population group. However, at the provincial level, depending on the standard errors obtained for these subclassifications, the estimates may not be statistically reliable for all cells and collapsing may be required.

5.2 Parameters to be estimated

The most important parameter to be estimated in a PES is the 'net census coverage error rate', universally known as the 'net omission rate', or the 'undercount'. It is based, in turn, on the 'dual-system estimate of the true population'. Next in importance is the calculation of adjustment factors to be applied to the census counts for the purpose of correcting for undercoverage or overcoverage. The adjustment factors are related to the undercount rates and are also based on the dual-system estimate of the true population. (See Section 8.3 for estimation methodology.)

5.3 Choice of procedure for coverage analysis

There are three alternative procedures for evaluating census coverage in a PES. These three procedures are known as A, B, and C (see Appendix I for definition of 'mover' terms).

Procedure A

- seeks to reconstruct the households as they existed at the time of the census
- the respondent must identify all persons in the sample household on the census reference date
- the aim is to match these persons (non-movers and out-movers) to the census questionnaires
- and to estimate the number and percentage of matched non-movers and movers (out-movers)
- the matching of non-movers and out-movers is relatively simple and inexpensive because the search is limited to sample areas
- but it is difficult and expensive to locate out-movers, especially out-mover households, given that they are no longer at the sample address (information when available is reported by proxy respondents)
- hence, there is a strong possibility of underestimation of the number of movers (out-movers).
- this leads to underestimation of the census omissions.

Procedure B

- seeks to identify all persons in the sample household on the reference date of the PES
- people respond for themselves; hence, field enumeration is more complete than in Procedure A
- the aim is to match these persons (non-movers and in-movers) to the corresponding census records
- and to estimate the number and percentage of matched non-movers and movers (in-movers)
- it provides a better estimate of the number of movers than procedure A
- but associated difficulties and costs of matching are far greater because it involves searching for in-movers in the area where they were enumerated during the census
- it is not sure if failure to match means an actually omitted person or an incorrectly located person
- this leads to overestimation of the census omissions.

Procedure C

- seeks to identify all persons in the sample household on the reference date of the PES and, in addition, any other persons in the household on the reference date of the census
- and to classify each person as either non-mover, out-mover, or in-mover with regard to his household presence status on the census date
- the aim is to match to the census records only the persons present on the date of the census, that is, the non-movers and the out-movers
- estimates for the number of non-movers and movers are based on in-movers (as in procedure B)
- matching rates for movers are estimated based on out-movers (as in procedure A).

In summary, Procedure C is a combination of Procedures A and B which takes advantage of the features of each to reduce matching difficulties and, at the same time, improve the estimation of movers. For this reason, the chosen procedure for PES 2001 was Procedure C.

6. SAMPLE PLAN

6.1 Sampling frame and sampling units

Since the PES methodology calls for a two-way match with census records, one criterion for the choice of a primary sampling unit is that the areas must have boundaries that are well-defined on geographic maps and recognizable on the ground. **The boundaries for the PES areas must correspond exactly to those for the census areas.** Another criterion for the choice of a primary sampling unit is, ideally, a size that is small (about 100 households) and uniform from area to area.

In order to meet these criteria, the best choice for the PES sampling unit was the census enumeration area (EA). The base for the sampling frame is the geographic information system (maps and database of EAs) developed for the census. For perfect independence between the PES and the census, the PES would ideally have used its own separate frame of EAs. In practice, however, this would have never been feasible given the enormity of cost, time, and resources necessary to develop a frame.

The PES sample is a single-stage cluster sample with the EA as the cluster. Once the EAs were selected, the final sampling unit was the household in housing units or hostels. All households and hostels in a sample EA were enumerated to permit matching against census records. (From the point of view of statistical efficiency, it is normally more efficient to take a subsample of households within each EA and spread the sample over more EAs to decrease the clustering effect for the same sample size, but this alternative makes matching impossible.)

6.2 The P sample and the E sample

The PES actually involves two samples, named the P sample and the E sample. The P sample or 'population' sample consists of the PES sample EAs drawn from the same target population, but independently from the census, for the purpose of estimating census omissions when compared to census records. The E sample is the 'enumeration' sample drawn from cases already enumerated in the census, but selected for independent checks for the purpose of estimating census erroneous inclusions when compared to original census records. Not all census-enumerated cases belong in the E sample: cases that are out of scope for the PES (for example, student residences and institutions) are not included in the E sample. The estimate of erroneous inclusions provides a correction factor needed in the dual-system estimate of the true population.

Even though theoretically the E sample may be separate from the P sample, in practice, it is better to allow it to overlap completely with the P sample to reduce costs and improve the precision of the estimates. The E sample then consists of the same EAs selected for the PES. A two-way match is conducted between the P sample and the E sample to identify both the omissions and the erroneous inclusions. The matching also produces the estimate of the 'matched population' component required in the dual-system estimator.

6.3 Stratification

To improve the efficiency of the PES sample design, the sampling frame was divided into homogeneous strata. For this purpose, variables correlated with coverage error were chosen, such as geographic area, since density and accessibility affect the quality of the census coverage. In addition, geographic stratification is necessary to obtain separate estimates by domain of analysis. Therefore, the first level of stratification corresponds to the geographic domains of estimation defined, namely province and urban/non-urban zones of residence.

For secondary stratification, advantage is taken of other variables correlated with the extent of coverage, such as subdivisions that are well delimited and possess a high degree of internal homogeneity with regard to socio-demographic characteristics. Hence, the sampling frame of EAs was substratified by geography type: urban formal, urban informal, tribal area, and rural formal.

To provide further stratification implicitly, the EAs were ordered geographically within each EA type and systematic selection was used.

6.4 Sample size and allocation to domains

The overall sample size for the PES was limited to 600 EAs due to operational and budget constraints. As mentioned, the domains of estimation or publication areas for which reliable estimates can be expected with this sample size are:

- National
- Urban/non-urban at national level
- Province

The allocation to the provinces and the expected standard errors for the census undercount rate are shown in Table 6.1 below. The absolute standard error was expected to be around 1 percentage point at the province level, except for the Northern Cape, where it was expected to be about 2 percentage points.

Table 6.1
Sample allocation to provinces and
expected standard errors for Census undercount rate

Province	Modified proportional allocation of 600 EAs	Expected SE for census undercount rate
Eastern Cape	90	0,0083
Free State	40	0,0108
Gauteng	100	0,0061
KwaZulu-Natal	100	0,0075
Limpopo	70	0,0101
Mpumalanga	50	0,0112
North West	50	0,0066
Northern Cape	40	0,0215
Western Cape	60	0,0076
South Africa	**600**	**0,0031**

Since the reliability of the PES estimates for individual provinces depends on the sample size allocated to that province, it was important to ensure a minimum number of sample EAs for the smallest provinces. This is reflected in the sample size for the smallest provinces which was increased to a minimum of 40 EAs.

Standard errors were calculated for the PES results using the CENVAR module of the US Census Bureau's CSPro/IMPS software, for the purpose of reporting on the reliability of the estimates as well as for planning future sample sizes. At the national level, a standard error of 0,5 percentage point (vs. the expected 0,3) was obtained. At the province level, the standard error was within 1 percentage point as expected, for all provinces but two. These two,

KwaZulu-Natal and Gauteng, achieved standard errors of 2,8 and 1,7 percentage points, respectively, largely due to the loss of sample EAs (Section 8.1.1). Absolute errors (1,96 times the standard error) for estimates of undercount rate are reported in Section 2.2. Future PES's will benefit from an increase in sample size from 600 to about 800 EAs.

6.5 Sample allocation within domains and sample selection

The following modified proportional allocation was adopted within domains. Table 6.2 shows explicit strata and substrata and the fixed number of sample EAs allocated to each.

Table 6.2 – Sample allocation within domains

	Urban	Non-urban	Total
Eastern Cape	34	56	90
Free State	24	16	40
Gauteng	84	16	100
Kwazulu-Natal	48	52	100
Limpopo	17	53	70
Mpumalanga	23	27	50
North West	21	29	50
Northern Cape	24	16	40
Western Cape	44	16	60
South Africa	319	281	600

Within each explicit substratum independently, the procedure was to select the EAs systematically with equal probabilities, after geographic ordering.

When there are good measures of size, sampling with probability proportional to size can be used to increase the efficiency of the sample design. However, in PES 2001, EA sizes were either nonexistent or not reliable. The selection with equal probabilities permitted a self-weighting sample in each of the explicit strata. Systematic selection offered convenience and efficiency, since it provided implicit stratification with the EAs ordered geographically.

The sample selection was carried out using the *SurveySelect* module from the SAS Software.

7. DATA COLLECTION, MATCHING AND PROCESSING METHODOLOGY

7.1 Summary of PES operational phases

The sequence of PES operations was the following.

- A pilot PES was conducted in March 2001 in 60 EAs in conjunction with the pilot census. This test permitted the evaluation and improvement of the PES questionnaire and procedures, and also provided feedback for the census procedures.

- The PES fieldwork – listing and enumeration – was conducted in 600 sample EAs across all nine provinces from 7 to 30 November 2001. All dwellings and structures within the boundaries of the selected EAs were listed, and households in housing units and hostels were interviewed using the PES questionnaire.

- The initial matching phase involved searching through the census records for the selected EAs in order to find the cases corresponding to the PES enumeration records, and vice versa.

- The reconciliation operation, consisting of field follow-up visits, also called 'control visits', followed the initial matching phase.

- The final matching phase used the results of the reconciliation visits to assign a definitive match status to each pending case.

- In the data capture phase, questionnaires were manually keyed into data files. Data entry was also verified.

- A data validation phase took place to detect and correct missing or otherwise invalid PES variables such as enumeration status and final match status, and to ensure data file integrity.

7.2 Questionnaire design

The PES questionnaire is presented in Appendix II. This questionnaire consists of some demographic questions from the census questionnaire as well as questions aimed at identifying the mover status of households and persons in the PES. The PES questionnaire was designed for automated data capture (by scanning) and semi-automated (computer-assisted) matching. However, due to problems with both the scanning system and the matching software, a manual processing system was implemented instead.

7.2.1 Including everyone

In accordance with Procedure C, the PES data collection aimed at identifying all persons in the household at the time of the census as well as those at the time of the PES. In a *de facto* enumeration this means persons who spent the reference nights in the household.

Hence, the goal was to enumerate every person – young or old, including babies, elderly persons, visitors, and non-citizens – who was present in the household on either or both reference nights, 6-7 November 2001 (PES night) and 9-10 October 2001 (census night). See Appendix I for definitions of 'absent' and 'present'.

The household list was established in Question P-00 and asked in two parts: (a) and (b), explained below. However, the names were required to be written as one continuous list, with the set (b) persons added immediately after those in set (a).

NAME (P-00) ASK:
a. Please tell me the names of all persons who spent the night between 6 and 7 November in this household. b. In addition, tell me the names of any persons who did not spend the night between 6 and 7 November, but who did spend the night between 9 and 10 October in this household.

The enumerator first listed, as the (a) set, all persons who were present in the household on the night between 6 and 7 November, that is, who spent the PES reference night in the household. These were the non-movers and the in-movers.

Then he added, as the (b) set, in addition to those present on PES night, all persons who although absent on PES night were present on census night (between 9 and 10 October), that is, who spent the census reference night in the household. These were the out-movers.

Rules of inclusion were the same as for the census:

- Babies born before midnight on the reference night and persons who died after midnight on the reference night were counted as present.
- Members of the household who were absent overnight, for example, working, travelling, or at an entertainment venue, were counted in the household if they returned to it the next day and did not stay at another place where they might have been enumerated.
- Paid domestic workers were counted as a separate household even if they lived in the same housing unit as the employer.

7.2.2 Determining presence for each person

With the use of Procedure C, the questionnaire must allow for the classification of each listed person as ‘non-mover’, ‘in-mover’, ‘out-mover’, or ‘out-of-scope’, with regard to their household presence status on census night.

The actual classification was to be done by computer during processing. To permit this, the questionnaire sought to ascertain the whereabouts of each person listed for <u>each</u> of the two reference nights, whether present in the household or not present (elsewhere, unborn or deceased).

PRESENCE ON PES NIGHT (P-02)	PRESENCE ON CENSUS NIGHT (P-03)
Based on P-00, write down where the person spent the night between 6 and 7 November. 1. In this household 2. Elsewhere 3. Deceased	Ask: Where did (person) spend the night between 9 and 10 October? 1. In this household 2. Elsewhere 3. Unborn

Even though the enumerator did not have to classify each person, what he in fact listed was:

- as part of set (a):
  - the non-movers – those who were present in the household on PES night and also present on census night;
  - the in-movers – those who were present in the household on PES night but were absent (elsewhere) on census night; and
  - those who were born after census night.

- and as part of set (b):
  - the out-movers – those who were absent (elsewhere) on PES night but present on census night; and
  - those who died after the census (also treated as out-movers).

7.2.3 Other PES-specific items

Fields for match status, census enumeration status, and the transcription of matched data were not included in the PES questionnaire because a computer-assisted matching operation had been planned and this information was to have been maintained by the computer system. Reconciliation Visit forms for follow-up were to be generated by computer following the initial matching. Instead, the RV forms were all handled manually. In future PES's, fields required for manual processing will be kept on the questionnaire even if automated processes are to be used.

7.2.4 Socio-demographic variables

Certain variables from the census questionnaire were repeated in the PES questionnaire for matching and content analysis purposes:

- Age
- Sex
- Relationship to head of household
- Marital status
- Population group
- Home language, and
- Highest level of education.

To ensure comparability between the PES and the census, the same wording, response categories and precodes, and also the same concept definitions, were maintained in the PES.

7.3 Fieldwork

7.3.1 General

A PES is not a mere repetition of the census, but a thorough enumeration in its own right. The goals of the PES exercise were to conduct an exhaustive enumeration, without omission or duplication, of all the households contained within the boundaries of the selected EAs. It meant not enumerating outside the EA boundaries. It also meant an exhaustive enumeration within each household and hostel of all persons present on PES night or census night, or both. Furthermore, it meant accurately measuring the characteristics of the enumerated population. At the same time, efforts were made to maintain operational independence between the PES and the census, as the validity of the PES estimates rests upon this fundamental assumption.

A PES must take place immediately after the census in order to minimise changes in the composition of households between the two dates. Still, enough time is needed to retrieve all the census materials from the field to avoid any contamination between the two operations. The census enumeration was scheduled to take place during the period of 10 to 31 October 2001 and PES 2001 was scheduled to take place from 7 to 22 November 2001, 4 weeks after the beginning of the census. The reference nights were PES night, the night of 6-7 November, and census night, the night of 9-10 October.

However, due to the extension of census field operations into November (and in a few areas into December), the PES fieldwork was delayed in some areas. In most areas, PES enumeration began on 15 November and ended on 7 December. Still, the reference night for the PES remained the night of 6-7 November.

Like the census, the PES was a *de facto* enumeration, which means people were enumerated based on their presence in the household on the reference nights, as opposed to their usual place of residence.

The supervisors and enumerators for the PES were drawn from the Stats SA household survey programme, to take advantage of their qualifications as experienced survey-takers and to ensure independence from the census. Detailed training guides were developed and training was held at head office for provincial managers and assistant managers, who then trained field staff in each province.

To further ensure independence from the census, efforts were made to guarantee that PES personnel did not have any preliminary knowledge of the census results for the areas for which they were responsible. At the same time, it was ensured that census field staff had no prior knowledge of the EAs that were going to be in sample for the PES.

7.3.2 Listing and enumeration

The PES fieldwork consisted of two phases:

- Listing and
- Enumeration (Interviewing)

Instructions for the PES fieldwork are in the PES Fieldworker's Manual and the PES Supervisor's Manual.

In every selected EA, fieldworkers were required to conduct the listing independently of any previous census listing and to list all housing units and other structures (including vacant buildings, businesses, schools, etc.) There were two reasons for such a comprehensive rule of inclusion:

- to guarantee completeness (exhaustiveness) of coverage. For example, housing units on institutional grounds or business premises, shacks erected overnight on vacant lots, or unoccupied or seasonal dwellings that became occupied during enumeration might not have been accounted for if left out during listing; and
- to provide a series of landmarks and reference points for use during enumeration and during subsequent revisits ('reconciliation' visits).

Listing results were recorded in the PES Fieldworker's Summary Book or '09 Book'. Except in farm areas and areas where distances were great, fieldworkers then swapped EAs so that they did not enumerate the EA they had listed.

The enumerator's tasks were to:

- identify all housing units and hostels within EA boundaries and all eligible persons in each household, without relying on census results;
- identify and add any dwellings not recorded by the lister in the PES 09 Book;
- interview all individuals at each housing unit and workers' hostel and complete the PES questionnaires correctly in accordance with the instructions in the manual; and
- ask for the sticker that was left by the census enumerator and paste it or write its number on the space provided on the PES questionnaire.

7.4 Initial matching phase

Coverage status is determined through case-by-case comparison of the PES independently-enumerated cases with the original census records. A two-way case-by-case matching is conducted of the two sources: the PES questionnaires and the census questionnaires.

Matching plays an integral role in the dual-system methodology:

- It provides an account of the persons included in both sources, and of the persons included in one source and excluded from the other, based on direct observation. (The PES does not simply rely on people reporting that they were or were not enumerated in the census.)
- It also enables the discovery and removal of erroneous inclusions (fabrications, duplications, out-of-scope, geographic misallocations) in either source.

The PES data processing plan called for the use of computer-assisted (automatic) matching. However, as mentioned, due to problems with the scanning system and the matching software and the need to conduct reconciliation visits in the field in a timely manner, a manual processing system was implemented instead. This resulted in manual matching and manually generated reconciliation visit forms, which posed some difficulties.

Before matching began, a PES Enumeration Status was assigned based on questions P-02 (presence on PES night) and P-03 (presence on census night) to classify each person as:

Figure 7.1
PES enumeration status

1. non-mover (present on PES night, and also on census night)
2. in-mover (present on PES night, but absent on census night)
3. born after census
4. out-mover (absent on PES night, but present on census night)
9. PES out-of-scope

The initial matching phase involved searching through the census records in order to find the corresponding cases from the PES enumeration records, and vice versa (a two-way match). In this manner, both the cases enumerated in the census but missed in the PES and those enumerated in the PES but missed in the census were identified. Specifications for the matching operation can be found in the Matching Reference Manual.

The initial matching phase produced seven types of results:

Figure .7.2
Initial match status

1.	matched
2.	possible match
3.	in PES not in census - definite non-match
4.	in PES not in census - insufficient or unclear information
5.	in PES not in census - in-mover
6.	in PES not in census - born after census
7.	in census not in PES

Cases of 'possible match', 'in PES not in census – insufficient or unclear information' and 'in census not in PES' were identified for reconciliation visits. In addition, reconciliation visits were carried out in all EAs with boundary problems or with overall quality results.

7.5 Reconciliation visits

The reconciliation operation consisted of field follow-up visits, also called 'control visits', to the PES sample EAs, following the initial matching phase. Reconciliation visits (RVs) are not considered optional, but are an integral part of the PES dual-system estimation methodology, since they provide follow-up for the E sample and help eliminate cases with insufficient information for matching.

The purpose of the reconciliation visits was to determine the final status of the unresolved cases identified above, specifically:

- to resolve the final match status for possible match cases;
- to determine whether households and persons enumerated in the census but not in the PES were correctly or erroneously enumerated in the census;
- to clarify doubtful cases or cases with insufficient or unclear information in order to assign a final match status; and
- to investigate EAs where boundary or enumeration quality problems were suspected.

The purpose of the reconciliation visits was not to add to the census enumeration or to the PES enumeration (since this would violate independence), but only to verify information already collected. To preserve the independence between the PES and the census, RV fieldworkers were not allowed to add persons or households, or change the recorded demographic characteristics of persons or households in either the census or the PES questionnaires.

Instructions for the reconciliation visits can be found in the manual titled 'Instructions for Reconciliation Visits'. The reconciliation visits gathered the following data:

(1) for possible matches:
- if a match was found, the matching person number
- if a match was not found, cases were treated as situation (2) below.

(2) for persons with a census record and no PES record as well as for persons with a PES record and no census record:
- presence on census night
- remarks to facilitate matching.

The RV forms also recorded whether the dwelling was found inside or outside EA boundaries, or not found at all. For census-enumerated cases, the RV fieldworker also noted whether or not the dwelling was seasonal. In all cases, an interview completion status was also recorded.

7.6 Final matching phase

The final matching phase used the results of the reconciliation visits to assign a definitive match status to each pending case.

The possible outcomes for the final matching phase are shown in the figure below:

Figure 7.3 – Final match status

1. matched

<u>In PES not in census</u>:
2. missed in census
3. PES erroneous inclusion - cases in PES not in census that were outside the EA boundaries or otherwise erroneously included in PES
4. PES insufficient information - cases in PES not in census for which a final match status cannot be assigned due to insufficient information
5. in-mover
6. born after census

<u>In census not in PES</u>:
7. correctly enumerated in census, missed in PES
8. Census erroneous inclusion
9. Census insufficient information – cases in census not in PES for which a final match status cannot be assigned due to insufficient information

Cases of 'possible match' were eliminated. All outcomes are mutually exclusive. For example, the category 'in PES not in Census: missed in Census' excludes other PES cases not in the census such as 'PES erroneous inclusion', 'PES insufficient information', 'in-mover', or 'born-after'.

The matching operation resulted in the classification of all PES-enumerated persons and census-enumerated persons in the sample EAs in specific categories. Cases without a valid match status were later sent back to manual verification to be resolved. When the final phase of matching was completed, every PES-enumerated case and every census-enumerated case in the sample EAs fell into one of the following specific mutually-exclusive categories:

a. for the PES-enumerated persons (the P sample):

	PES eumeration status	Final match status
❑ matched non-mover	1	1
❑ matched out-mover	4	1
❑ non-matched non-mover	1	2
❑ non-matched out-mover	4	2
❑ in-mover	2	5
❑ born after	3	6
❑ PES erroneous inclusion	1 through 9	3
❑ PES insufficient information	1 or 4	4

b. for the census-enumerated persons (the E sample):

	Final match status
❑ matched	1
❑ Census correctly enumerated (non-matched)	7
❑ Census erroneous inclusion (non-matched)	8
❑ Census insufficient information	9

At this point the questionnaires proceeded to the data capture phase.

7.7 Data capture and data validation

As mentioned, the PES data processing plan called for automated capture by scanning. For this reason, the questionnaire and reconciliation visit forms were designed for automated processes. Due to problems with the scanning system and the need to conduct the field reconciliation visits in a timely manner, a manual processing system was implemented instead. This required, in addition to manual matching, key-from-paper data entry and manual processes for tracking.

The PES data entry application was designed using the US Census Bureau's CSPro CSEntry module.

The switch to manual processing with a questionnaire not designed for this purpose caused many inefficiencies in the data entry system, and contributed to significant delays and quality

problems. To remedy these and other quality problems, a data validation phase was implemented.

The PES key-from-paper entry was inefficient because the PES questionnaire was not designed for manual processing. Therefore, rather than 100% verification, dependent verification was used, whereby clerks visually inspected captured questionnaires on screen and compared each entry with the paper questionnaires.

Initially, verification was set up in two parts: 100% verification of problematic or suspicious cases – which referred to the fact that all the questionnaires (rather than a sample thereof) flagged as having problems were subject to data entry verification – and sample verification of normal questionnaires. However, as new edit flags were developed, most questionnaires ended up going through data entry verification.

Once a questionnaire was selected for review, data entry was verified for all of its captured fields. The verifier:

- searched for the questionnaire in the database,
- compared the screen version of the questionnaire with its paper version,
- examined all captured fields, and
- corrected any errors.

Questionnaires were flagged for review if any of the following applied to any person in the questionnaire:

- final match status indicating match but no matching record found
- final match status missing or otherwise invalid
- PES enumeration status missing or otherwise invalid
- PES enumeration status incompatible with final match status
- incorrect questionnaire barcodes,
- incorrect person numbers, and
- miscellaneous others.

Aside from the above, in 64 EAs the integrity of the data file became suspect. These files were abandoned and a completely fresh data entry operation was carried out with 100% independent rekeying.

In total, the percentage of questionnaires subjected to data entry verification amounted to approximately 90%.

The validation operation also served to resolve by automatic correction incompatibilities between the PES Enumeration Status and the Final Match Status, when these were found even after the manual verification. It must be noted that the scope of PES edits was limited to checks on file and record integrity, structure checks, and errors in PES variables such as match status, enumeration status, person number, and related information. In a PES, socio-demographic data are not edited, since one of the PES purposes is to evaluate variability between PES responses and census responses.

In spite of the computer edits and manual verification, it was suspected that erroneous inclusions still persisted. Consequently, a field 're-visit' exercise was set up whereby the worst 66 EAs (1,1% of the EAs) in terms of match rates were identified for boundary checks. The re-visits revealed that about two-thirds of these EAs in fact had boundary interpretation problems and contained erroneous inclusions either in the PES or in the census, or in both. This exercise led to additional corrections in the data files.

8 ESTIMATION PROCEDURES

8.1 Sampling weights

8.1.1 Base sampling weights

The sampling frame consisted of 72 487 EAs (after deleting the vacant, institution, recreational and industrial EAs) from the Geographic Information System database of August 2001. The sample allocation is indicated in Section 6.5.

The PES sample was a one-stage cluster sample and the PES EAs were drawn with equal probability within explicit strata. An equal-probability selection method was used with systematic selection of ordered EAs to draw a sample in each explicit stratum, using the SAS procedure *suveyselect*.

Consequently the base sampling weight of a sample EA was equal to the reciprocal of its probability of selection, that is, to the universe total number of EAs in the stratum divided by the number of sample EAs in that stratum. Within each EA, the weight for each household and each person was equal to the EA sampling weight, since their probability of selection, given the selection of the EA, was equal to one.

8.1.1 Adjustments to base sampling weights

Due to an inadvertent over-stratification of the sampling frame (by including 'EA type'), the sampling weights in the sample added to 72 029 instead of 72 487. This distortion was corrected using a marginal benchmarking programme with the correct explicit stratification marginal totals (number of EAs) contained in the sampling frame. With the marginal benchmarking, the sampling weights added to the correct total of 72 487.

A certain amount of substitution was necessary during the fieldwork. Some PES sample areas became known to census workers before they had completed their work. To avoid contamination, the entire sample of EAs was re-selected in these provinces. The new sample of EAs was selected in the office by PES technical staff using the same probability-sampling method as the original sample. Since all this happened before PES enumeration began, this is not an actual case of substitution. However, in five other cases, EAs were replaced after enumeration began. Two vacant EAs were replaced to maintain the same sample size. In another occasion, a vehicle hijacking caused three EA boxes of completed questionnaires to be lost, and these EAs were also replaced in the sample. The EAs were substituted with adjacent EAs (rather than randomly selected Eas)to avoid rearranging supervisory areas. Because the substitute EAs were of the same substratum as the original EAs, their weights were not adjusted.

Weighting adjustments were also necessary to account for the loss of EAs in the sample. A total of 13 sample EAs were dropped for the following reasons:

- In one case, the PES EA box was completely lost.
- In another case, the EA consisted entirely of out-of-scope living quarters.
- In 11 cases, the PES missed the sample EA altogether and went to a different area instead. In three of these 11 cases, the PES fieldworker visited undefined areas with no recognisable geographical boundaries. In the other eight, the area enumerated corresponded to another EA which was not in the sample. Upon examining the proportion of the EA enumerated, it was found in all cases, that the EA was not covered in its entirety.

The decision to drop EAs was made in order to reduce the bias the missed EAs would introduce in the coverage estimates. The disadvantage, however, is a reduction in the precision (standard error) of the estimates, particularly at the provincial level.

For all the dropped in-scope EAs, the remaining EAs in the stratum were reweighted. In the case of the out-of-scope EA, the original weight was maintained for the other EAs in the stratum to avoid overestimating the in-scope population. This reduced the sum of the weights to 72 371, the estimated in-scope population.

8.2 Coverage evaluation: Calculation of dual-system estimates for persons

Coverage measures were calculated only for cases belonging to the universe of interest (Section 1.2).

1. The initial estimates – weighted estimates of total from the sample – using Procedure C, are (also see Figures 8.1 and 8.2):

    a. total number of non-movers in the universe (P sample);
    b. total number of out-movers in the universe (P sample);
    c. total number of in-movers in the universe (P sample);
    d. total number of matched non-movers in the universe (P sample);
    e. total number of matched out-movers in the universe (P sample);
    f. total number of matched in-movers in the universe;
    Note: in Procedure C, the number of matched in-movers cannot be calculated directly, given that no match is attempted for the in-movers in the sample. But the 'out-movers' and the 'in-movers' constitute the same group in the universe: the 'movers', assuming a closed population. Therefore, an assumption can be made that, in the universe, the match rate for in-movers would be the same as that for out-movers (estimated by e/b). Hence, the total number of matched in-movers in the universe is estimated indirectly by [(e/b)*c].
    g. total number of census erroneous inclusions in the population (E sample);
    h. total number of cases correctly enumerated in the census but missed in the PES (E sample).
    i. total number and percentage of census persons with insufficient information (E sample).
    j. total number and percentage of PES erroneous inclusions and PES insufficient information cases (P sample).

Figure 8.1
Initial derivations in dual system estimation

	Parameter	Derivation
I1	Estimated no. of non-movers and % of total population represented by non-movers	NM / PES Pop
I2	Estimated no. of out-movers and % of total population represented by out-movers	OM / PES Pop
I3	Estimated no. of in-movers and % of total population represented by in-movers	IM / PES Pop
I4	Estimated no. and rate of matched non-movers	Matched NM / NM
I5	Estimated no. and rate of matched out-movers	Matched OM / OM
I6	Estimated no. of matched in-movers	I5 rate * I3 total
I7	Estimated no. of census erroneous inclusions	weighted sum
I8	Estimated no. of census correctly enumerated persons missed in PES	weighted sum
I9	Estimated no. and % of census persons with insufficient information	divide by A1a
I10	Estimated no. and % of PES erro incl and PES insuff info cases	divide by A2a

2. The 'matched' population is given by the total number of matched non-movers plus the estimated total number of matched in-movers in the universe.

MATCHED POP = MATCHED NON-MOVERS + ESTIMATED MATCHED IN-MOVERS

Figure 8.2
Analysis derivations in dual system estimation

	Parameter	Derivation
A1a	Census population (uncorrected for err incl and insuff info)	(I4 + I6) + I7 + I8 + I9
A1b	Census population (corrected for err incl and insuff info)	(I4 + I6) + I8
A2a	PES population	I1 + I3
A2b	Matched population	I4 + I6
A3	PES persons missed in Census – Total	A2a – (I4 + I6)
	PES persons missed in Census – Rate	divided by A2a
	Coverage rate	[1 – A3 rate]
A4	Census correctly enumerated missed in PES	I8
	Census correctly enumerated missed in PES – Rate	divided by A1b
A5	Census erroneous inclusions – Total	I7
	Census erroneous inclusions – Rate	divided by A1a
A6	Preliminary dual-system est. of true pop	(A1b * A2a) / Matched pop
A7	Net error (net undercount) – Total	A6 – A1a
	Net error (net undercount) – Rate	divided by A6
A8	Gross error – Total	A3 rate * A6
	Gross error - Rate relative to true pop	A3 rate
A9	'Adjustment factor' for Census	A6 / A1a
	Final dual-system estimate of true pop	A9* census count

3. The E-sample estimate of the population enumerated in the census *[Uncorrected CENS_POP]* is the sum of the matched population, the population erroneously included in the census, the population correctly enumerated in the census but missed in the PES, and the census insufficient-information cases.

$$\text{CENS\_POP\_UNCORR} = \text{MATCHED\_POP} + \text{CORR\_ENUM} + \text{ERR\_INCL} + \text{INSUFF\_INFO}$$

The census population corrected for erroneous inclusions and insufficient-information cases *[Corrected CENS_POP]* is calculated without adding these last two categories.

$$\text{CENS\_POP\_CORR} = \text{MATCHED} + \text{CORR\_ENUM}$$

4. The P-sample estimate of the total population [*PES_POP*] is the sum of the non-movers and in-movers in the population.

$$\text{PES\_POP} = \text{NON-MOVERS} + \text{IN-MOVERS}$$

5. The PES-enumerated population missed in the census is calculated by subtracting the matched population from the PES estimate of the total population to obtain:

$$\text{PES\_POP\_MISSED\_IN\_CENSUS} = \text{PES\_POP} - \text{MATCHED\_POP}$$

The rate of PES population missed in the census is the missed population above relative to the PES estimate of total population.

6. The estimated total number of census erroneous inclusions *ERR_INCL* is the same as calculated in the initial tables. It includes fabrications, duplications, and geographic misallocations, etc. As mentioned, the main purpose of the E sample is to provide an estimate for this variable in order to permit a correction in the dual-system estimate of the true population.

The census erroneous inclusion rate is equal to the total number of persons erroneously included in the census relative to the E-sample estimate of the census population.

7. The preliminary dual-system estimate of the 'true' population *[TRUE_POP]* is the population estimated from one source (the PES) multiplied by the population estimated from the other source (the census, after correcting for erroneous inclusions and insufficient information) and divided by the population found in both:

$$TRUE\_POP = \frac{PES\_POP \times Corrected\ CENS\_POP}{MATCHED}$$

8. The net coverage error – universally known as the 'net omission', or the 'undercount' – is the difference between what should have been counted (true population) and what was counted (census population). The net coverage error represents the undercount still remaining in the census figures even after the partial cancellation caused by the overcount.

$$\textit{Net Undercount} = \text{TRUE\_POP} - \text{CENS\_POP\_UNCORR}$$

The net coverage error rate – the 'net omission rate' or the rate of 'undercount' – is the total net error relative to the dual-system estimate of the true population, that is, divided by TRUE_POP. **This measure constitutes the single most important indicator of the quality of the census coverage.**

9. The gross coverage error – the 'gross omission' – is, as defined in this context, what the census truly missed without taking into account the overcount. It is the gross omission relative to the true population, as opposed to the net omission, that is, without being offset by the erroneous inclusions. It corresponds to the estimation method used in the South African 1996 PES.

$$\begin{aligned} \textit{Gross Coverage Error} &= \textit{Population Found in PES Missed in census} \\ &\quad + \textit{Population Missed in Both census and PES} \end{aligned}$$

$$= PESPersonsMissedinCensus + \frac{(TruePop - CensPopCorr) \times (TruePop - PESPop)}{TruePop}$$

$$\textit{Gross Error Rate} = \frac{\text{Gross Error}}{\text{True Population}}$$

Equivalently:

$$\begin{aligned} \textit{Gross Error Rate} &= \textit{1 – Matched Pop/PES Pop} \\ &= \textit{rate of PES persons missed in census (Table A3)} \end{aligned}$$

which means the total gross error can be calculated as:

$$\textit{Gross Error Total} = \textit{Rate PES persons missed in census} \times \textit{True Pop}$$

10. The final dual-system estimate of the *True Population*, which corresponds to the 'Adjusted Population', is obtained through the use of a ratio estimator of total, which is superior in accuracy to the preliminary estimate, by reducing both variance and bias.

$$\left[\frac{Preliminary\, TRUE\_POP}{CENS\_POP\_UNCORR} \right] * \textit{Actual Census Count}$$

where the ratio inside the bracket represents the 'adjustment factor' for the census count.

11. The relation between the undercount rate and the adjustment factor is the following:

$$ADJ\_FACT = \frac{1}{1 - undercount\ rate}$$

In other words, the adjustment factor is the reciprocal of the complement of the undercount rate.

For example, an undercount rate of 2% implies an adjustment factor of 1,0204. Likewise, an undercount rate of 8% implies an adjustment factor of 1,0870, and an undercount rate of 14% implies an adjustment factor of 1,1628, and so forth.

12. Another way of viewing the adjustment factor is the following:

$$\text{Adjustment Factor} = \frac{PES\_POP \times CENS\_POP\_CORR}{MATCHED\_POP} \Big/ CENS\_POP\_UNCORR$$

If we consider the quantity $\frac{MATCHED\_POP}{PES\_POP}$ as the 'Coverage Rate', then:

$$\text{Adjustment Factor} = \left[\frac{1}{Cov\ Rate}\right] \times \left[\frac{CENS\_POP\_CORR}{CENS\_POP\_UNCORR}\right]$$

While the quantity inside the first bracket is clearly a correction for underenumeration, the quantity in the second bracket – which is the proportion of the census population that was correctly enumerated, i.e, not erroneously included – serves as a correction for overenumeration. Note that the South African 1996 PES adjustment factor is based only on the first quantity (see Section 10).

Hence, the final adjusted population is in effect calculated as follows:

Adjusted Population = *underenumeration correction factor* $\times$ *overenumeration correction factor* $\times$ *census count*

Also note that the underenumeration correction factor is always ≥ 1 and the overenumeration correction factor is always ≤ 1. The overall factor can theoretically fall on either side of 1, depending on which is higher, the undercount or the overcount.

13. The probabilities of inclusion and omission of a person are calculated as follows:

Figure 8.3
Derivation of probabilities of inclusion

P (included in Census)	=	Census Population Corrected / True Population
P (included in PES)	=	PES Population / True Population
P (included in both Census and PES)	=	P (included in Census) * P (included in PES) *per independence assumption*
P (included in Census, but missed in PES)	=	P (included in Census) * [1 - P (included in PES)]
P (included in PES, but missed in Census)	=	P (included in PES) * [1 - P (included in Census)]
P (missed in both Census and PES)	=	[1 - P (included in Census)] * [1 - P (included in PES)] *per independence assumption*

14. The distribution of the true population – based on the preliminary dual-system estimate, after removing erroneous inclusions and insufficient-information cases in census – is the following:

Figure 8.4
Derivation of population distribution estimates

		Census Enumeration		
		Included	**Omitted**	**Total**
PES Population	**Included**	MATCHED POP	in PES, missed in Census	PES POP
	Omitted	in Census, missed in PES	missed in both	
	Total	CENSUS POP CORR	GROSS OMISSION	TRUE POP

It is given by:

Census pop corrected for err incl & insuff. info	=	P(included in census)	× dual-sys estimate of pop
PES pop (excludes err. incl. & insuff. info)	=	P(included in PES)	× dual-sys estimate of pop
Pop included in both census and PES	=	P(included in both census and PES)	× dual-sys estimate of pop
Pop included in census, missed in PES	=	P(included in census, but missed in PES)	× dual-sys estimate of pop
Pop included in PES, missed in census	=	P(included in PES, but missed in census)	× dual-sys estimate of pop
Pop missed in both census and PES	=	P(missed in both census and PES)	× dual-sys estimate of pop

8.3 Coverage evaluation for households

A working definition for households first had to be established. A PES 'household' was defined as a parent questionnaire (the one containing the head of household) without the continuation questionnaires. (Continuation questionnaires were excluded in the PES household analysis due to linking errors; it is estimated that 1,9% of households had more than 10 persons and thus required continuation questionnaires). By this definition, 'questionnaire' and 'household' refer to the same set of persons. The definition of household was driven primarily by the P-sample. The total number of matched households was calculated as the total number of matched questionnaires from the P-sample. If one P-sample questionnaire matched two E-sample questionnaires, it was counted only once; if two P-sample questionnaires matched to one E-sample questionnaire, each P-sample questionnaire was counted separately.

In the processing of the census itself, the household was a parent questionnaire combined with any continuation questionnaires for it. A post-capture processing operation was performed to establish the proper links. Even though the basic definition for household is similar in both the census and the PES, there will be conceptual differences because the 'questionnaire' is not a fixed entity in the universe: the number of questionnaires completed for one housing unit can vary from interview to interview. A fixed entity would have been the housing unit, as identified by an address or physical structure. But PES did not do address matching and, in any case, the census did not define households by address, as in South Africa there are many cases of households sharing addresses (or dwellings), so they still would not have agreed.

Next, the 'PES enumeration status' and 'match status' were defined for each household. If at least one person in the questionnaire was matched, then the household was considered matched. If all persons in the questionnaire were missed, then the household was considered as a miss (in the census or in the PES). For the balance of questionnaires, priority was given to 'erroneous inclusion' and then to 'insufficient information'. Hence, if at least one person in the questionnaire was erroneously included, the household was considered as erroneously included (in the census or in the PES). Otherwise, the household was considered an 'insufficient information' case (in the census or in the PES).

Once these variables were defined, the same dual-system estimation procedure defined in Section 8.3 for persons was applied.

8.4 Formation of adjustment classes

The overall coverage estimates when broken down to geographic or demographic variables (such as province, sex, age group or population group) could be skewed due to the fact that persons and households are not evenly missed over such subgroups of the population. Homogeneous adjustment classes, i.e., classes within which coverage rates are more or less the same, are thus formed and a single adjustment factor is then calculated in each of the adjustment classes independently. The adjustment classes were obtained by using the Automatic Interaction Detection (AID) technique XAID (cf. Hawkins, D.M. 'FIRM – Formal Inference-based Recursive Modeling', Release 2.0, Technical Report No. 546, University of Minnesota, USA, 1990). There are two AID techniques, CHAID and XAID, the former being applicable in the case where the dependent variable is a dichotomous variable and the latter where the dependent variable is a continuous variable.

A matching variable was created which took on the value 1 for a non-mover person (or household) if the person was counted in the census, the value 0 if the person was missed in the census and, for in-movers, a continuous value between 0 and 1. The latter occurrence necessitated the use of XAID. The matching variable, in this case, can be interpreted as a probability that the person (or household) was enumerated in the census.

For persons, the predictors used (per province) were: geography type, sex, age group, and population group. For households, the predictors used were: province, geography type, size of the household, and population group of the head of the household.

The XAID model determined combinations of the predictors that were statistically significant in modeling the coverage probability. The characteristics defined by the XAID branches (i.e., the different branches in the dendrogram created by XAID) were then taken as the adjustment

classes. The ‘stopping rules’ used in the XAID person runs were: a minimum of 1000 cases for a group to be analysed and a maximum of 50 groups for splitting. Furthermore, raw as well as Bonferoni significance levels of 1% for splitting/merging were specified. The same stopping rules were used for households, with the only exception being that the minimum number of cases for a group to be analysed was set as 2000.

After the creation of these various adjustment classes, a separate adjustment factor was calculated for each class, using the formulas described in Section 8.3. Due to the fact that the factors are ratios, the population when adjusted at the national or at the provincial level is not equal to the summation of the adjusted population over all adjustment classes. This is an inherent mathematical inequality – a difference between totals produced using combined ratios vs. separate ratios – and not a calculation error.

One issue was whether the national adjusted population should be the separate ratio estimate of total (summing up the adjusted population across adjustment classes) or the combined ratio estimate of total using the national adjustment factor. The separate-ratio estimate produced a lower variance (because of homogeneous classes with high heterogeneity among them, with a sufficient number of observations in each class) but it has a higher bias than the combined-ratio estimate. The combined-ratio estimate had higher variance but its bias is lower than that of the separate-ratio estimate, due to the consistency property of ratio estimators (which makes the bias diminish as n gets larger). Nevertheless, since each class had a large number of observations, the separate-ratio estimate was chosen.

As a result of this ‘bottom-up’ approach, the undercount rate was re-calculated in each publication cell as:

Adjusted in-scope population – Unadjusted in-scope population

8.5 Application of adjustment factors to census data

The adjusted population is obtained by multiplying the appropriate adjustment factor to the actual census count in the adjustment class, and then summing across classes. In practice, this is equivalent to using a standard weighting procedure where the ‘weight’ corresponds to the adjustment factor.

As mentioned in the discussion of ‘PES target universe’ in Section 1.2, the PES was limited to a large subset of the population. Because the coverage rates in the balance of population are unknown, no adjustment was made for these persons.

Hence, as a first step in the application procedure, the total universe for the census was partitioned into two sets: ‘Population within in-scope subuniverse’ and ‘Balance of population’. Each person or household was first determined to be in or out of the target population based on EA type, living-quarters type, and questionnaire type. Only eligible cases, i.e., cases in the in-scope subuniverse, received the designated adjustment factors. Non-eligible cases, i.e., balance-of-population cases, received an adjustment factor of 1.

The eligible person was then assigned on an individual level the adjustment factor corresponding to the adjustment class he belonged to, according to province, geography type, sex, age group, and population group. Similarly, each household was assigned on an individual level the adjustment factor corresponding to the adjustment class it belonged to,

according to province, geography type, size of the household, and population group of the head of the household.

Census counts, both unadjusted and adjusted, were then calculated separately for the two population subsets:

Unadjusted census population =
Unadjusted 'Population within in-scope subuniverse'
\+ *Unadjusted 'Balance of population'.*

Adjusted census population =
Adjusted 'Population within in-scope subuniverse'
\+ *Unadjusted 'Balance of population'.*

It is worth nothing that PES adjustment factors were based on the original geographic and demographic classifications of persons. For geography type and EA type, 'original' referred to the classification in the August 2001 frame, before EA type changes occurred. For living-quarters type and for demographic variables, 'original' refers to these variables as originally reported in the census.

Therefore, to maintain compatibility between the distribution of PES cases and census cases, the **original** classifications (i.e., unedited or 'raw' data before editing and imputation) were used to decide which factor a person would receive. Thus, census persons received the adjustment factor corresponding to their original geography type and EA type, original living-quarters type, and original sex/age group/population group cell. Once the adjustment factors were applied, persons and households were permitted to shift to post-editing classification cells (which render census data more accurate and more meaningful), but they carried their original adjustment factors individually into their new cells.

8.6 Content evaluation for persons

Content analysis is discussed in Section 4.1. The following must be noted regarding the use of the PES for the measurement of content error:

- It is limited to matched cases.
- It is limited to the in-scope subuniverse, consisting of housing units and hostels within in-scope EA types.
- The PES is not assumed to provide the 'truth'; therefore, response bias is not measured, only response variance.
- Comparison is of unedited PES and census sociodemographic responses. (PES sociodemographic data are not subject to edit; census data are, but these edits take place outside the PES.)
- Unlike the census and PES questionnaires in the PES sample, data capture for the full census was not by key-from-paper but by scanning with rigorous quality control. In addition, census data were later subject to an intensive edit and automatic-correction process. Hence, to a certain extent, the data quality in the published census results is improved over what is indicated by the content analysis.

It was also noted in Section 4.1 that the estimated person totals shown in the content analysis tables do not coincide with the final census totals for each characteristic because:

- they are based on the sample of census records in the PES and are, therefore, subject to sampling variability;
- they include only matched cases, not the full sample;
- they are unedited while the census characteristics are edited;
- they include only the in-scope subuniverse while the final census totals include the full universe; and
- they are unadjusted while the final census totals are adjusted.

The sole purpose of these totals is to compare the census responses with the PES responses and to calculate the measures of consistency; they are not for socio-demographic analytical purposes.

It is further noted that, although the content tables were supposed to include all matched persons, about 11% of the person records did not have their matching companion because of barcode/person number errors. They were thus omitted from the content error calculations. To the extent that these 11% might reflect different consistency patterns, the content error measures might be somewhat biased.

Variability is measured by means of four different indicators: the net difference rate, the index of inconsistency (simple and aggregate), the gross difference rate, and the rate of agreement. Appendix III provides an illustration of the computations for the net difference rate, the index of inconsistency, and their standard errors and confidence intervals. Source: 'Evaluating Censuses of Population and Housing', pages 87-91, Statistical Training Document, ISP-TR-5, U.S. Census Bureau, 1985.

8.6.1 Net difference rate (NDR)

The net difference rate is the difference between the number of cases in the census and the number of cases in the PES that fall under each response category, relative to the total number of matched persons in all response categories. The NDR formula for the i-th category is:

$$NDR = \frac{Y_{\bullet i} - Y_{i\bullet}}{n} \times 100$$

$$for\ i = 1, ..., C$$

where: $Y_{\cdot i}$ = *unweighted census number of cases in i-th category*
$Y_{i\cdot}$ = *unweighted PES number of cases in i-th category*
n = *unweighted number of matched cases*
C = *total number of response categories for characteristic 'y'*

8.6.2 Index of inconsistency

The index of inconsistency is the relative number of cases for which the response varied between the census and the PES. It is the ratio of the simple response variance to the total variance of the characteristic, including its variability in the population.

It is calculated for each response category 'i' according to the following formula:

$$\hat{I} = \frac{(Y_{\bullet i} + Y_{i\bullet} - Y_{ii})}{\frac{1}{n}[Y_{.i}(n - Y_{i\bullet}) + Y_{i\bullet}(n - Y_{.i})]} \times 100$$

$$(i = 1,...,C)$$

where: Y_{ii} = *number of cases where category i was given as a response in both the census and the PES*

The following formula is used to calculate the **aggregate** index of inconsistency (that is, for all the response categories of the characteristic as a whole):

$$\hat{I}_{AG} = \frac{(n - \sum_{i}^{C} Y_{ii})}{(n - \frac{1}{n}\sum_{i}^{C} Y_{.i} Y_{i\bullet})} \times 100$$

8.6.3 Gross difference rate (also off-diagonal proportion)

The gross difference rate (GDR) is calculated for the characteristic as a whole. It is the number of discrepancies between the census responses and the PES responses relative to the total number of persons matched. It is equivalent to the sum of all cells off the diagonal, for all categories, or the complement of the sum of the diagonal cells.

$$GDR = \frac{(n - \sum_{i}^{c} Y_{ii})}{n} \times 100$$

8.6.4 Rate of agreement

The rate of agreement is the complement of the gross difference rate. A low rate of agreement indicates a high degree of variability, and viceversa.

$$Rate\ of\ Agreement = \frac{\sum_{i}^{c} Y_{ii}}{n} \times 100$$

9 ACCURACY OF DUAL-SYSTEM ESTIMATES

Estimates from a dual system are subject to certain assumptions and are affected by certain types of errors. The design of the PES included measures to control these different errors.

9.1 Assumptions of the dual-system method

The fundamental assumptions of the dual-system estimation method are: a closed population, independence between the census and the PES, no erroneous inclusions in the census or the PES, and no incomplete matches. To the extent that these assumptions are violated, bias is introduced in the estimates. Biases are difficult to measure. However, steps are taken during the design and implementation of the PES to ensure that the source and effect of these biases are controlled.

a. Closed population

   A closed population is one whose composition remains relatively unchanged over the time between the two studies (the census and the PES); this means an insignificant number of external migrations. This is why it was important to conduct the PES as soon as possible after the census. In PES 2001, the number of external migrations during the interim period between the two operations was assumed to be small.

b. Independence between the census and the PES

   The validity of the dual-system estimates is based upon a fundamental assumption of independence between the PES and the census. Ideally, it would involve a separate frame for the EAs, different enumerators, separate processing, etc. In practice, it is never possible to obtain absolute independence. However, it is still necessary to separate the two operations to the extent feasible.

   Hence, in PES 2001, certain measures were taken to maintain operational independence, such as: (1) having PES technical staff reporting to a separate unit in the organisation; (2) separating administrative, logistical, and managerial structures in the field; (3) waiting until the census enumerators left the EA before conducting the PES enumeration (census workers went back to some EAs after the PES had taken place, but whether or not an area was revisited by census had nothing to do with its being in the PES); (4) ensuring that PES areas were covered by field staff other than census enumerators, namely Stats SA household survey staff; (5) ensuring that PES personnel had no preliminary knowledge of the census results for the EAs they were responsible for; (6) likewise, ensuring that census personnel had no preliminary knowledge of the areas that would later be in the PES; and (7) keeping the PES processing separate from the census processing.

c. No erroneous inclusions in either system

   Erroneous inclusions (also called overenumeration) consist of duplications, fabrications, geographic misallocations (related to boundaries), and the inclusion of persons not belonging to the target population.

Under the original dual-system model, applied under ideal conditions, the census population total and the PES population total (in the numerator of the estimation formula for *TRUE_POP*) would be free from erroneous inclusions. However, in practice, there are erroneous enumerations and, consequently, the model must be applied under these realistic conditions. It is necessary, therefore, to identify and remove the erroneous inclusions from the totals.

In the case of the P sample, adequate quality control and follow-up allowed the identification and removal of these cases from the sample itself. Hence, they did not have a chance of coming into the PES estimate of the total population. However, in the case of the census total, the erroneous enumerations are already included in the census count, hence overestimating this population. Since these erroneous inclusions will not be found in the matched population (in the denominator of the *TRUE_POP* formula), the PES estimate of the census total is corrected to remove them, in order not to overestimate *TRUE_POP*. The use of the E sample aimed at identifying the different types of erroneous enumerations in the census total. They were detected through the two-way match and the reconciliation visits (and boundary check re-visits).

The estimation formula for *TRUE_POP* allows the subtraction of erroneous inclusions from the census total to correct for their effect.

d. No incomplete matches

As in the preceding case, the dual-system model does not take into account cases included in the census total or the PES total (in the numerator of the estimation formula for *TRUE_POP*) which can never be found in the matched population (in the denominator) because they lack sufficient information to be matched. Theoretically, any failure to match should be due to actual omission and not to the inability to match.

In practice, there may be cases with insufficient information to complete the matching. Fortunately, with the implementation of the reconciliation visits after the initial matching phase, a final match status was resolved for the vast majority of cases. The remaining number of insufficient-information cases was close to 1,6% for the E sample (census cases) and to 1,1% for the P sample (PES cases).

In the case of the PES population, these cases were removed from the P sample itself and did not go into the estimate. In the case of the census total, they would already be included in the census count, and it was thus necessary to correct the PES estimate of the census total to subtract them, as in the case of erroneous inclusions, to avoid overestimating *TRUE_POP*.

e. No assumption regarding which system is better

Sometimes claims are made that PES estimates are closer to the 'true' value than census estimates. This is not one of the assumptions of the dual-system estimation model: the dual system only provides an estimate of the cases included in one source (the PES) and excluded from the other (the census), and vice versa. Both estimates contribute to the dual-system estimate, which is more complete than either the census or the PES estimate alone.

9.2 Errors affecting the dual-system estimates

The total error of dual-system estimates, as with any other sample estimate, includes sampling error (variance mostly and bias) and non-sampling error (bias mostly and variance). The most relevant types of non-sampling bias – non-response bias, correlation bias, and matching bias – are discussed below. Sampling variance and sampling bias are also discussed.

a. Non-response bias

Non-response occurs in every census or sample survey. Non-response is technically different from non-coverage in that the non-respondents are accounted for during listing but no questionnaires are processed for them. It results from refusals, non-contacts, and unusable questionnaires. A thorough and vigorous effort was made to follow up the non-respondents and complete the interviews.

Some of the factors that affected the PES response were: negative reaction to census enumeration (that had already taken place in the area), a mistaken perception on the part of the public that the 'census operation was over' and that further enumeration was not legitimate, response burden (having to complete a census interview plus a PES interview), and proximity to the December holiday season.

b. Correlation bias

Correlation bias is the tendency of cases included in the census to have a higher probability of inclusion in the PES than cases not included in the census. One reason is that the same persons tend to be missed in both the PES and the census because they are members of population subgroups which are difficult to cover. Good quality control during the PES is essential to improve the enumeration, especially for these hard-to-enumerate cases, in order to reduce the correlation bias. Another reason for correlation bias is lack of operational independence between the PES and the census. The correlation bias is lower when in-movers rather than out-movers are enumerated.

c. Matching bias

This type of bias refers to an error in the matching process, which occurs in two forms: erroneous matches and erroneous non-matches. It is necessary to minimise both types of matching error. The matching bias is usually lower when out-movers rather than in-movers are matched.

Problems that can affect match rates include: boundary interpretation problems, misplacement of census questionnaires or PES questionnaires, lack of diligence in reconciliation visits, insufficient operational control, and data capture difficulties. These problems were alleviated through remedial efforts.

d. Variance

As with all estimates obtained through a sampling procedure, PES estimates are subject to sampling variability. Estimates of the variance (standard error) were calculated for PES

2001 parameter estimates. Absolute errors and confidence intervals are reported for the undercount rate and the adjusted population in Part I of this report.

e. Sampling bias

Many of the estimators used in the dual-system procedure are ratio types, which are biased but consistent. Their bias approaches zero as the effective sample size (upon which the estimate is based) increases. Hence, efforts were made to maintain an adequate sample size in each estimation cell.

f. Substitution bias

Since substitution of EAs took place in the PES after the sample was selected (a total of five sample EAs were replaced, using adjacent EAs), bias may have been introduced in the estimates. Bias is inherent in any substitution, but the extent of bias is dependent on the rate of substitution and on the differences between the original EAs and the new EAs with regard to census coverage rate. There is no reason to believe great differences existed since adjacent EAs within a substratum tend to be very similar. Not only is the substitution bias considered minimal and insufficient to distort the coverage estimates, but the use of substitution actually served to prevent more serious errors (increase in variance due to sample loss and high refusal rates), which would have lessened the accuracy of the coverage estimates.

10 COMPARISON OF 1996 AND 2001 ADJUSTMENT METHODS

The following is an assessment of the major differences in methodology between the 1996 and the 2001 South African post-enumeration surveys with regard to various aspects.

10.1 Dual-system estimation

A dual-system estimation method involves the preparation of estimates based on matching cases from two different and independent sources describing the same events. The matching permits an estimate of the number of cases reported by one source but omitted in the other.

Both PES 2001 and PES 1996 used the dual-system methodology. However, PES 2001 used an additional ‘enumeration’ sample (see Section 10.2 below) to provide corrections in the estimation process; PES 1996 did not. PES 2001 involved a two-way match and PES 1996, a one-way match. In PES 2001, census *erroneous inclusions* were identified and removed from the true population total; in the 1996 total, these cases were included. Both PES 2001 and PES 1996 identified and removed PES erroneous inclusions (out-of-scope cases).

The following diagrams help illustrate the differences:

Coverage distribution of census enumeration (uncorrected for erroneous inclusions)

	Census enumeration
Total excluding erroneous inclusions	A+C
Included in PES	A
Omitted from PES	C
Census erroneous inclusions	E
Total including erroneous inclusions	A+C+E

Coverage distribution of true population based on dual system

		Census enumeration (corrected for erroneous inclusions E)		
		Included	Omitted	Total
PES Enumeration	Included	A	B	A+B
	Omitted	C	D	C+D
	Total	A+C	B+D	A+B+C+D

PES 2001 identified each of components A through C plus another component, the census erroneous inclusions (E), separately. Component D was not observed directly but was accounted for mathematically. It then removed E from the census population which, when corrected, stands at A+C. The true population is equivalent to A+B+C+D.

PES 1996 did not identify C separately, but took it into account indirectly when it estimated the census population, which consisted of A+C+E. Component D was also accounted for mathematically even though it was not observed. However, the 1996 method did not identify and remove E from the census population, hence overstating the true population as A+B+C+D+E. As a consequence, the net undercount was also overestimated.

10.2 Use of the E sample

As mentioned, PES 2001 involved two samples – the P sample and the E sample – while PES 1996 involved only one, the P sample. The 'population' sample or P sample consists of the PES sample cases drawn from the target population for the purpose of estimating cases omitted in the census. The E sample is the 'enumeration' sample drawn from the cases already enumerated in the census, for the purpose of estimating census erroneous inclusions. The P and E samples are discussed in greater detail in Section 6.2.

10.3 Reconciliation visits

PES 2001 included reconciliation visits; PES 1996 did not. These control visits carried out after the initial matching phase not only help resolve doubtful cases but they provide needed follow-up for the E sample. They are the vehicle for identifying census erroneous inclusions (component E) and census correct enumerations missed in the PES (component C). They also help minimise insufficient information cases.

10.4 Comparison of formulas

$$\text{1996 estimate of true population} = \frac{Census\ Pop \times PES\ Pop}{Matched\ Pop}$$

where, Matched Pop = PES persons found in census

$$\text{1996 adjustment factor} = \frac{PES\ Pop}{Matched\ Pop}$$

$$\text{2001 estimate of true population} = \frac{(Census\ Pop - ErrIncl) \times PES\ Pop}{Matched\ Pop}$$

where, Err Incl = census erroneous inclusions

$$\text{2001 adjustment factor} = \frac{(Census\ Pop - ErrIncl) \times PES\ Pop}{Matched\ Pop} \Big/ Census\ Pop$$

[In actuality, a preliminary estimate of the true population is calculated where census Pop is an estimate based on the E sample. The adjustment factor is based on that estimate. Later when the full census count is produced, the adjustment factor is applied to it and the final estimate of the true population is calculated.]

Comparing the two adjustment factors, we see that:

$$\text{2001 adjustment factor} = \frac{(Census\ Pop - ErrIncl) \times 1996\ Adj\ Fact}{Census\ Pop}$$

$$= (1 - ErrIncl\ Rate) \times 1996\ Adj\ Fact$$

where Err Incl Rate = census Erroneous Inclusions / Census Population

The quantity in parentheses is always less than 1 (or equal to 1 if there is no overcount). It acts as a 'correction' for census overenumeration. From this relation, it is clear that the 2001 adjustment factor by definition will always be smaller than the 1996 factor for the same data.

If the 2001 net undercount rate had been calculated using the 1996 formula, it would have been equal to 20,0% instead of 17,6%. Conversely, a rough approximation to what the 1996 net undercount rate would have been if the 2001 formulas had been used is 8,29% instead of 10,69% (if we assume that the same erroneous inclusion rate of 2,4% also occurred in 1996).

10.5 Treatment of movers

PES 2001 used Procedure C while PES 1996 used Procedure B (see explanation of these coverage analysis procedures in Section 5.3). The 2001 PES procedure is less susceptible to matching error than the 1996 PES procedure (Procedure C is expected to yield a lower matching bias than Procedure B). As a result, overestimation of the non-match rate and thus of census omissions was more likely in 1996 than in 2001. Therefore, this aspect as well as the estimation formula itself would have contributed to a higher adjustment factor in 1996.

10.6 Other differences

The analysis presented assumes all other factors in the two implementations to be constant. Yet, this is not the case. Many more cases were classified as insufficient information in 1996 (22% of P-sample cases) than in 2001 (1,6% of E-sample cases and 1,1% of P-sample cases). Many cases that were classified as sure misses in the 2001 procedures might have been classified as unresolved in the 1996 procedures. In 1996, the 22% 'unresolved' cases underwent a modeling process which assigned probabilities of coverage based on the distribution of resolved cases. A logical conclusion from these facts is that the 1996 PES was more likely to underestimate the census omissions than the 2001 PES.

10.7 Conclusion

The PES 2001 methodology was much more comprehensive than the 1996 methodology. The estimation formulas and the treatment of movers for the 1996 PES might have led it to overestimate the net undercount rate. At the same time, the lack of resolution for unresolved cases might have led it to underestimate it. On net balance, it is difficult to conclude which of the two scenarios dominated and, therefore, whether the 1996 adjusted population was too high or too low.

Appendix I – Relevant PES definitions

Enumeration

Enumeration is the process of counting all the members of a defined population and collecting demographic and other information about each person. This counting takes place by means of administering a questionnaire to all households.

Enumeration area

An enumeration area (EA) is the smallest geographical unit (piece of land) into which the country is divided for census enumeration purposes. Each EA is expected to have clearly defined boundaries. EAs typically contain between 100 and 250 households.

In the PES, the EA serves as a sampling unit, that is, an area which can be selected to be in sample. Selected EAs have to be completely enumerated by PES Fieldworkers during the allocated time.

De facto enumeration

A *de facto* enumeration is one in which people are enumerated according to where they stay on the reference night, not according to usual place of residence (*de jure*). The South African Population Census and the PES are *de facto* enumerations.

Household

A household is a group of people who live together and provide themselves jointly with food or other essentials for living, or a single person who lives alone. Since this is a *de facto* enumeration, only people present in the household on the reference nights are included as part of the household.

A household is not necessarily the same as a family.

Household head

In the first instance, the head of household is the person that the household regards as such. If necessary, the head can be defined as the main decision-maker, or alternatively, the person who owns or rents the dwelling, or the person who is the main breadwinner. The head can be either male or female.

If two people are equal decision-makers, the older of the two should be named as head of the household. In a household of totally unrelated persons, the oldest should be named as the household head.

Dwelling

A house, tent, hut, houseboat, etc. where one or more households live. A dwelling may be constructed or converted for human habitation or not intended for habitation but actually used for such purpose at the time of the PES.

Unoccupied dwelling

Premises built specially for living purposes, which are suitable for occupation, but which are not occupied during the PES, for example, an empty house or an empty flat in a block of flats.

Seasonal dwelling Dwellings usually occupied only at certain times of the year which remain unoccupied the rest of the year, such as holiday homes, harvest-time homes, etc. These types of dwelling must be identified as such in the 09 Book (Column 5) and on the PES questionnaire (Question H-05).

Housing unit A unit of accommodation for a household, which may consist of one structure, or more than one structure, or part of a structure. (Examples of each are a house, a group of huts, and a flat.) It may be vacant, or occupied by one or more than one household.

A housing unit has a separate entrance from outside or from a common space, as in a block of flats.

Premises not intended for use as living quarters, but used for human habitation, such as a barn, warehouse, etc., are also classified as housing units for census and PES purposes.

NB The term housing unit is contrasted with collective living quarters – i.e. all living quarters are either housing units or collective living quarters.

Collective living quarters Living quarters where certain facilities are shared by groups of individuals or households. They can be divided into: (a) hotels, motels, guesthouses, etc. (b) workers' hostels and student residences; and (c) institutions.

Listing The process of identifying and recording all housing units and all other structures in each EA. The list is compiled in the 09 Book. Listing instructions are provided in Part III of this Manual.

Non-response Non-response is the absence of interview data for a household identified in the listing. It results from refusals, non-contacts, unusable questionnaires, etc. A thorough and vigorous effort must be made to minimise these sources, as non-response introduces serious bias in the survey results.

Non-contact Non-contact describes the situation where a fieldworker fails to:

- locate or reach a dwelling because he's run out of time or resources, or because there are errors in the 09 Book, or
- make contact with a household because no-one was at home at the time of the visit. Repeat visits are necessary to try to find someone at home.

Refusal A refusal occurs when the fieldworker fails to gain cooperation from a household once contact has been made. In these circumstances the fieldworker fills in a refusal form. Every effort is then made by senior PES officials to persuade the household to complete the interview.

Census night Census night is the night between **09 and 10 October 2001**. It is the reference date of the census and it is referred to in the PES questionnaire. Census night is one of the two reference dates which are the basis for the inclusion of an individual in the PES questionnaire, and for the identification of movers between the time of the census and the time of the PES.

PES night PES night is the night between **06 and 07 November 2001**. It is one of the two reference dates which are the basis for the inclusion of an individual in the PES questionnaire, and for the identification of movers between the time of the census and the time of the PES.

Non-movers Persons who were present in the household on the night between 6 and 7 November, that is, the reference night for the PES, and who were also present on the night between 9 and 10 October, that is, the reference night for the census, including babies, the elderly, visitors, and non-citizens.

In-movers Persons who were present in the household on the night between 6 and 7 November, that is, the reference night for the PES, but who were absent on the night between 9 and 10 October, that is, the reference night for the census, including babies, the elderly, visitors, and non-citizens.

Out-movers Persons who were absent from the household on the night between 6 and 7 November, that is, the reference night for the PES, but who were present on the night between 9 and 10 October, that is, the reference night for the census, including babies, the elderly, visitors, and non-citizens.

Born after the census Babies who were present in the household on the night between 6 and 7 November, that is, the reference night for the PES, but who were not yet born as of the night between 9 and 10 October, that is, the reference night for the census. Even though these babies are included in the list of household members, they are different from the in-movers, because they are out of the scope of the target population, that is, the population as of census night.

Present vs. absent 'Present' means the person spent the reference night (for the census or for the PES) in the household. Presence always refers to the reference night and not to the date when the enumerator (Census or PES) comes to interview. Persons to be counted as 'present' include regular family members who spent the reference night in the household, as well as any visitors, non-relatives, or non-citizens who also spent the reference night in the household. All these types of persons also include babies and elderly persons.

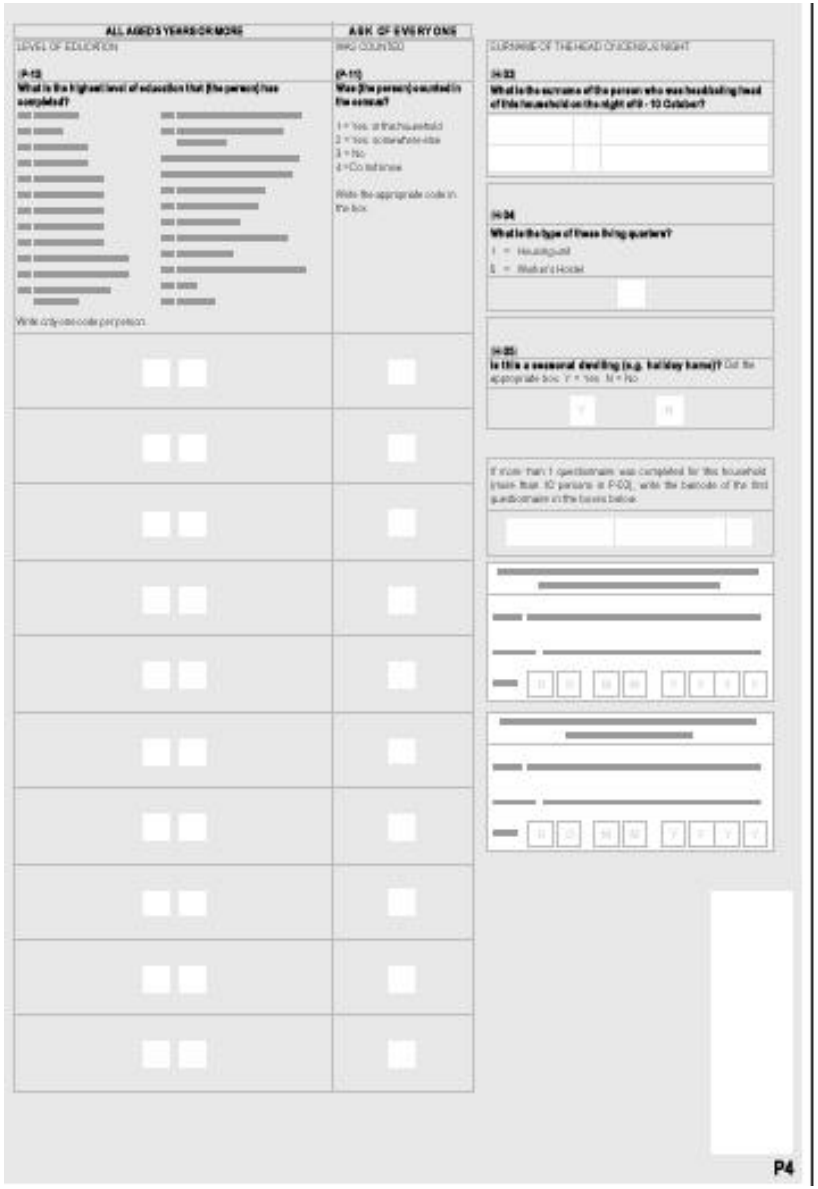
'Absent' means the person spent the reference night (for the census or for the PES) elsewhere and not in the household being interviewed. As with presence, it is with regard to the reference night and not to the

date when the enumerator (Census or PES) comes to interview. Persons to be considered as ‘absent’ include even regular family members, including babies and elderly persons, if they did not spend the reference night in the household.

Note that, for each reference night, members of the household who are away overnight, for example, working, travelling, or at an entertainment venue, and did not stay in another household or place where they might be enumerated, are to be counted as present in the household if they return to it the next day.

In addition, babies born before midnight of each reference night and persons who died after midnight of each reference night are to be counted as present for that reference night.

Appendix II PES questionnaire



ALL AGED 5 YEARS OR MORE

LEVEL OF EDUCATION

P-10

What is the highest level of education that [the person] has completed?

Write only one code per person.

ASK OF EVERYONE

RKG COUNTED

P-11

Was [the person] counted in the census?

1 = Yes, at this household
2 = Yes, somewhere else
3 = No
4 = Could not say

Write the appropriate code in the box.

SURNAME OF THE HEAD OF DE FACTO HOUSEHOLD

H-03

What is the surname of the person who was the de facto head of this household on the night of 9 - 10 October?

H-04

What is the type of these living quarters?

1 = Housing unit
2 = Worker's Hostel

H-05

Is this a seasonal dwelling (e.g. holiday home)? Cut the appropriate box. Y = Yes, N = No

Y N

If more than 1 questionnaire was completed for this household (more than 10 persons in P-03), write the barcode of the first questionnaire in the boxes below.

P4

PES

CENSUS 2001

POST ENUMERATION SURVEY 2001

FOR STATISTICAL USE ONLY

FOR OFFICE USE

EA number:

Record number:

Household number:

Province:

Local municipality:

Main place:

Sub-place:

Physical address:

Postal code:

Telephone no:

How many questionnaires were completed for this household?

If more than one questionnaire, what is the number of this questionnaire?

Number of people: Total Male Female

The PES questionnaire should be completed for all households living in housing units (including housing units on institutional grounds) and for all households living in hostels for workers.

How to fill in the questionnaire

Write the codes in the appropriate boxes or dot the box.

Use a pencil. If you make a mistake, use a soft rubber to erase wrong dots, numbers or letters and write the correct answer.

Please write dots, numbers and letters clearly so that the scanning machine can read them. Always use CAPITAL LETTERS.

For example, for a person who speaks English most often, write code 0 2 in the box in column P-08.

Make sure that all dots, numbers and letters stay inside the box.

(Correct) (Incorrect)

For example

1 2 3 4 5
6 7 8 9 0

Do not leave spaces or use hyphens when a word runs across more than one line. For example, Cape Town should be written in this way

C A P E T O
W N

Y = Yes, N = No

Y N

Date: D D M M Y Y Y Y

Place sticker here

P1

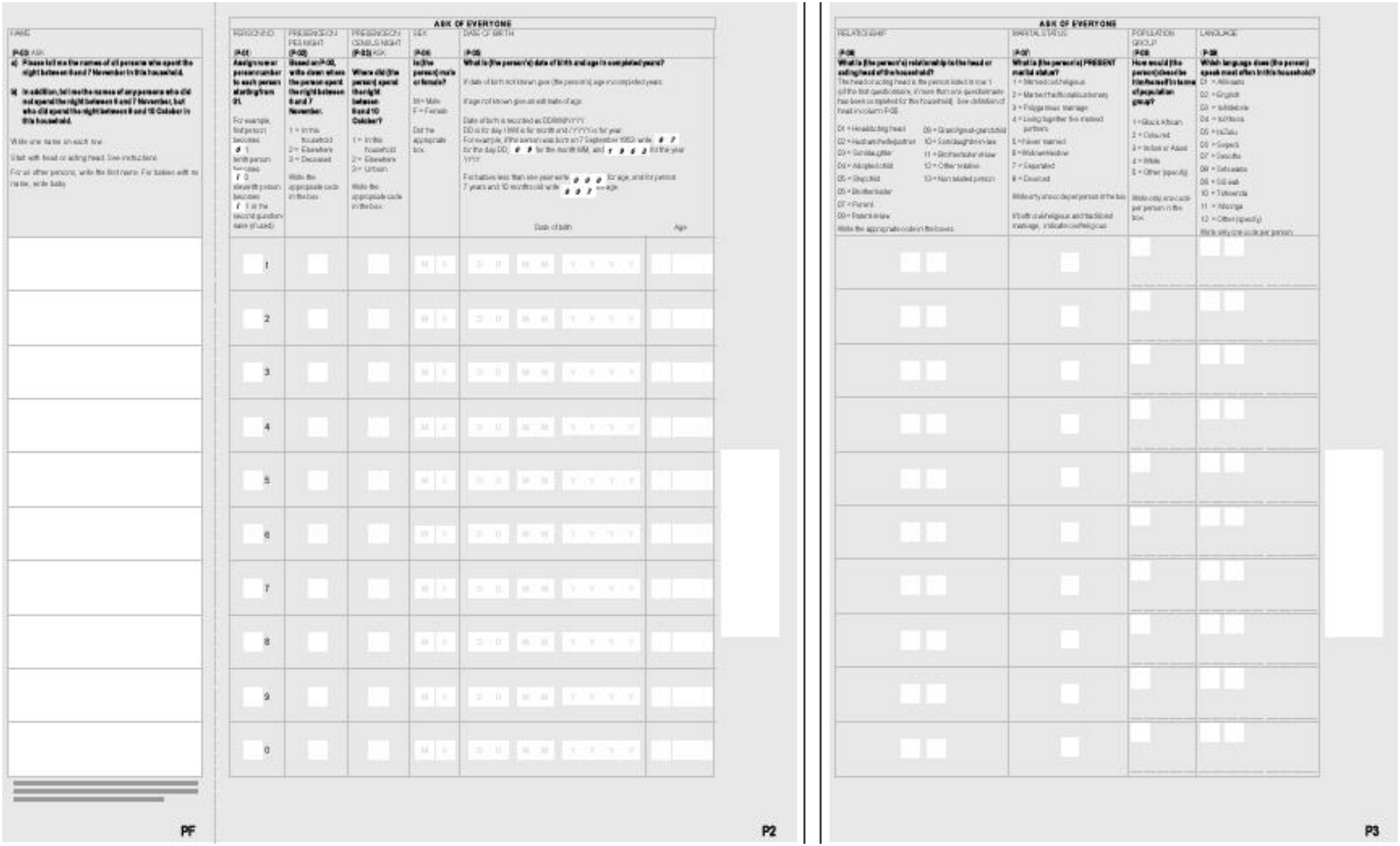
CONFIDENTIALITY

STATISTICS ACT NO.6, 1999

10(1) Despite any other law, no return or other information collected by Statistics South Africa for the purposes of official or other statistics that relates to an individual or a household may be disclosed to any person.

10(3) Any person who is involved in the collection of, or who may have, that information or data, must first take an oath of confidentiality.

19(a) & 19(g) Any officer of Statistics South Africa who willfully discloses any data or information obtained in the course of such employment to a person not authorised to receive that information is guilty of an offence and liable on conviction to a fine not exceeding R10 000, or to imprisonment for a period not exceeding six months or to both such fine and such imprisonment.


ASK OF EVERYONE
NAME
(P-03) ASK
a) Please tell me the names of all persons who spent the night between 6 and 7 November in this household.
b) In addition, tell me the names of any persons who did not spend the night between 6 and 7 November, but who did spend the night between 8 and 10 October in this household.
Write one name on each line.
Start with head or acting head. See instructions.
For all other persons, write the first name. For babies with no name, write baby.
RESPONDENT (P-01)
PRESENCE ON PES NIGHT (P-02)
PRESENCE ON CENSUS NIGHT (P-03) ASK
SEX (P-04)
DATE OF BIRTH (P-05)
What is (the person's) date of birth and age in completed years?
Date of birth
Age
RELATIONSHIP (P-06)
What is (the person's) relationship to the head or acting head of the household?
MARITAL STATUS (P-07)
What is (the person's) PRESENT marital status?
POPULATION GROUP (P-08)
How would (the person) describe the race/population group?
LANGUAGE (P-09)
Which language does (the person) speak most often in this household?
PF
P2
P3

Appendix III

Illustration of computations for net difference rate, index of inconsistency, standard errors and confidence intervals

Figure III.1
GENERAL NOTATION FOR COMPUTING RESPONSE ERROR MEASURES IN PES STUDIES

PES classification *(i=1,2,…..c)*	Total reporting	Classification reported in Census *(j= 1,2, …..c)*									
		Category *1*	Category *2*			Category *j*				Category *c*	
Total reporting*	**N**	$\mathbf{Y_{.1}}$	$\mathbf{Y_{.2}}$			$\mathbf{Y_{.j}}$				$\mathbf{Y_{.c}}$	
Category *1*	$\mathbf{Y_{1.}}$	Y_{11}	Y_{12}			Y_{1j}				Y_{1c}	
Category *2*	$\mathbf{Y_{2.}}$	Y_{21}	Y_{22}			Y_{2j}				Y_{2c}	
Category *i*	$\mathbf{Y_{i.}}$	Y_{i1}	Y_{i2}			Y_{ij}				Y_{ic}	
Category *c*	$\mathbf{Y_{c.}}$	Y_{c1}	Y_{c2}			Y_{cj}				Y_{cc}	

* *This table excludes all cases for which there was no report in either the census, the PES, or both*

Figure III.2
COMPUTING NET DIFFERENCE RATE AND INDEX OF INCONSISTENCY

Net difference rate for category *i* (an estimator of β (response bias) only when the PES response is considered to be the 'truth'):

$$NDR = \frac{(Y_{.i} - Y_{i.})}{n} \times (100), (i = 1,.....,c)$$

Index of inconsistency for category *i* (appropriate only when the PES response is considered to be in replication of the census response):

$$\hat{I} = \frac{(Y_{.i} + Y_{i.} - 2Y_{ii})}{\frac{(Y_{.i}(n - Y_{i.}) + Y_{i.}(n - Y_{.i})}{n}} \times (100), (i = 1,.....,c)$$

Note: Y_{ii} *is the* i^{th} *diagonal term*

Figure III.3
COMPUTING STANDARD ERRORS AND NINETY-FIVE PER CENT CONFIDENCE INTERVALS

Ninety-five percent confidence interval of net difference rate for category *i*:
(i=1,…,c)
Ninety-five percent confidence limits are:

$$\frac{(Y_{.i} - Y_{i.}) \pm 2\sqrt{Y.i + Yi. - 2Yii + 1}}{n} \times (100)$$

Exceptions:

(1) If $(Y_{i.} - Y_{ii}) = 0$, then widen the high ninety-five per cent confidence limit by adding: $\left[\frac{2}{n} \times (100)\right]$

(2) If $(Y_{.i} - Y_{ii}) = 0$, then widen the low ninety-five per cent confidence limit by subtracting: $\left[\frac{2}{n} \times (100)\right]$

(3) If both (1) and (2) above, the ninety-five per cent confidence limits are estimated as: $\left[\frac{-4}{n} \times (100)\right]$ to $\left[\frac{+4}{n} \times (100)\right]$

Ninety-five per cent confidence interval of index of inconsistency for category i: $(i=1,\ldots,c)$

(1) If $\left(\frac{Y_{.i} + Y_{i.} + 2Y_{ii}}{n}\right) \leq 0.10$, ninety-five per cent confidence limits are:

$$\frac{(Y_{.i} + Y_{i.} - 2Y_{ii} + 2) \pm 2\sqrt{(Y_{.i} + Y_{i.} - 2Y_{ii} + 1)}}{Y_{.i}\left(1 - \frac{Y_{i.}}{n}\right) + Y_{i.}\left(1 - \frac{Y_{.i}}{n}\right)} \times (100)$$

(2) If $\left(\frac{Y_{.i} + Y_{i.} + 2Y_{ii}}{n}\right) > 0.10$, ninety-five per cent confidence limits are:

$$\frac{(Y_{.i} + Y_{i.} - 2Y_{ii} + 2) \pm 2\sqrt{\frac{1}{n}(Y_{.i} + Y_{i.} - 2Y_{ii} + 1)(n_{..} - Y_{.i} - Y_{i.} + 2Y_{ii})}}{Y_{.i}\left(1 - \frac{Y_{i.}}{n}\right) + Y_{i.}\left(1 - \frac{Y_{.i}}{n}\right)} \times (100)$$

Ninety-five per cent confidence interval for the aggregate index of inconsistency:

(1) If $\left[\frac{n - \sum_{i=1}^{c} Y_{ii}}{n}\right] \leq 0.10$, ninety-five per cent confidence limits are:

$$\frac{\left(n - \sum_{i=1}^{c} Y_{ii} + 2\right) \pm 2\sqrt{n - \sum_{i=1}^{c} Y_{ii} + 1}}{\left(n - \frac{1}{n}\sum_{i=1}^{c} Y_{.i}Y_{i.}\right)} \times (100)$$

(2) If $\left[\frac{n - \sum_{i=1}^{c} Y_{ii}}{n}\right] > 0.10$, ninety-five per cent confidence limits are:

$$\frac{\left(n - \sum_{i=1}^{c} Y_{ii} + 2\right) \pm 2\sqrt{\frac{1}{n}\left(n - \sum_{i=1}^{c} Y_{ii}\right)\left(\sum_{i=1}^{c} Y_{ii}\right)}}{\left(n - \frac{1}{n}\sum_{i=1}^{c} Y_{.i}Y_{i.}\right)} \times (100)$$