Centre for Actuarial Research (CARe)

A Research Unit of the University of Cape Town

Estimation of fertility from the 2001 South Africa Census data

Tom A Moultrie
Rob Dorrington

Centre for Actuarial Research
University of Cape Town

CARe Monograph No. 12

ISBN:0-7992-2280-1

Centre for Actuarial Research
University of Cape Town
Private Bag
7701 Rondebosch
SOUTH AFRICA

+27 (0)21 650 2475 (T)
+27 (0)21 689 7580 (F)
care@commerce.uct.ac.za (E)

Centre for Actuarial Research (CARe)

A Research Unit of the University of Cape Town

Estimation of fertility from the 2001 South Africa Census data

Tom A Moultrie
Rob Dorrington

Centre for Actuarial Research
University of Cape Town

CARe Monograph No.12

Summary

Fertility in South Africa has been falling for almost four decades. The 2001 South Africa Census offers the opportunity to reflect on this decline, and to assess the trajectory and patterns of fertility in the country, among its population groups and in its provinces.

Analysis of the data in the 2001 census shows that fertility among all four main population groups continues to fall, and that the national level of fertility is now below three children per woman. The rate of decline indicated by the estimated levels of fertility is a continuation of the long trend of gradually declining fertility. The 2001 census was the second conducted in a post-apartheid South Africa. The first, which was conducted in 1996, is regarded as the most reliable and accurate enumeration of the South African population since that in 1970. The comparison of fertility levels and trends estimated from these two post-apartheid censuses provides valuable checks and comparisons that further enhance our understanding of fertility dynamics in the country.

Summary results, comparing the levels of fertility from the two most recent censuses, are shown in the table below.

Estimates of fertility levels by population group and province, 1996 and 2001 censuses

	1996	*2001*
National	**3.23**	**2.84**
African	3.49	3.04
Coloured	2.64	2.41
Indian/Asian	2.45	1.98
White	2.02	1.82
Western Cape	2.60	2.39
Eastern Cape	3.93	3.28
Northern Cape	2.80	2.43
Free State	2.97	2.53
KwaZulu-Natal	3.48	3.04
North-West	3.05	2.77
Gauteng	2.55	2.43
Mpumalanga	3.50	3.13
Limpopo	3.94	3.62

Nationally, fertility has fallen by 0.4 of a child per woman over a five year period, a decline of 12.1 per cent. The rate of decline has been fastest among Indian/Asian women, and slowest among Coloured women.

The level of, and the rate of change in, provincial fertility reflects both the composition of each province by population group as well as differences in the level of urbanisation. Fertility is lowest in the Western Cape, Northern Cape, Gauteng and the Free State. These four provinces also showed the lowest fertility levels in the 1996 census.

A particularly large decline in fertility is apparent in the data for the Eastern Cape. Whether this fall is 'real' or – more probably – a result of errors in the data collected in that province is unknown at this stage. Further research, based on household survey data, together with the Demographic and Health Survey (DHS) that is currently in the field, may be able to shed light on this question.

The derivation of robust and reliable fertility rates is a worthwhile endeavour in its own right and represents the major portion of the present work, but to do so requires that this report discusses the quality of the data collected in the 2001 census at some length.

We identify (and correct for) several significant anomalies in the data collected in the 2001 census. The two errors that give rise to the greatest concern are, first, the apparent inability of the census to capture accurately all births that occurred in the preceding twelve months. As with the 1996 census data, only about half of the 1.1 million births we estimate to have occurred in the country over the year before the 2001 census were actually enumerated. Errors of this sort are common, and well documented, in all developing country settings. Nonetheless, strenuous efforts should be made to improve the quality of recent fertility data collected in future censuses.

Second, the data on the number of children born to women of reproductive age are seriously deficient. All indications are that this was a result of inadequate (or incorrect) training of enumerators. This matter can, and must, be addressed before the next census if these data are to be useful. While a method is presented here that allows estimates of women's lifetime fertility to be calculated, the ramifications of this flaw in the data extend beyond the analysis and determination of fertility levels and trends. Most importantly, the data on the proportion of children born that are still alive indicates that (most probably) women interpreted the question on children ever born as asking about children still alive. This compromises the estimation of child mortality levels from the census data, since the estimation procedures rely heavily on those proportions. Further investigation of these data is outside the scope of this report, but is covered in a second monograph prepared by the Centre for Actuarial Research for Statistics South Africa (Dorrington, Moultrie and Timæus 2004).

Authors' Note

This report was prepared for Statistics South Africa in terms of a contract awarded by them in October 2003 to prepare an analysis of the fertility data collected in the 2001 census. The authors are grateful to Statistics South Africa for this opportunity, as well as for all their assistance in making the data available to us.

However, all errors are our own and all opinions are those of the authors and do not necessarily reflect those held by Statistics South Africa.

Tom A Moultrie and Rob Dorrington
Cape Town, December 2004

Table of Contents

List of Tables

List of Figures

1 Introduction

The second census since the first democratic elections in South Africa was conducted in 2001, exactly five years after the previous census. After several decades of uncertainty as to the exact nature of South African fertility levels and trends, data from these two censuses together with the 1998 South Africa Demographic and Health Survey (DHS) allow a detailed picture of recent South African fertility to be derived.

This report presents a detailed analysis of the fertility data collected in the 2001 census. The data, and the editing rules used to clean the 'raw' responses of the enumerated population, are subjected to close scrutiny, before being used to derive fertility rates by population group and province (for African South Africans and nationally). The institutional legacy of apartheid continues to affect the vital demographic statistics of the South African population in a way that is highly correlated with population group. Hence, our collective vision of a united non-racial South Africa notwithstanding, the measurement of demographic variables must still be stratified by population group. In addition, the collection and analysis of demographic indicators by population group assists in the monitoring of the country's progress in redressing historical iniquities and imbalances. The presentation in this report of demographic data by population group in no way condones or supports apartheid-era classifications.

As the second census since 1994, the data collected in 2001 can, and should, be compared with those from the 1996 census to better understand the dynamics of demographic change in the country. The data from the 1996 South Africa Census have been analysed extensively and reported on elsewhere by Moultrie and Timæus (2002; 2003). The age-specific and total fertility rates for the year centred on October 1996 as indicated by the 1996 census and the 1998 South Africa DHS by population group are shown in Table 1.1. These data show that, by developing country standards, fertility in South Africa was low (3.2 children per woman nationally, according to the DHS), and indeed it had been falling gradually for the better part of thirty years. Analysis of the data on birth histories collected in the 1998 DHS further suggest that the average time between births had almost doubled in the same period, from around 30 months to almost 60 months. Both these features of the South African fertility decline, which set it apart from almost all other declines, have been analysed in depth using a combination of demographic techniques, statistical modelling and historical and political analysis (Moultrie 2002).

While the levels of fertility indicated by the 1996 census and the 1998 DHS do not correspond that well, Moultrie and Timæus (2002, 2003) have shown that the age distributions of fertility indicated by each data source are fundamentally similar. In their conclusion, Moultrie and Timæus (2003) suggest that differences and errors in sample design, data collection (the DHS tended to reflect a more urbanised and better educated population than that indicated by the census), and weighting may account for the differing estimates of the level of fertility, and suggest (contrary to the norm) that the estimates derived from the census should be used in preference over the results from the DHS.

Table 1.1 Age-specific and total fertility rates in 1996 by population group and nationally, 1996 census and 1998 DHS

	Africans		*Coloureds*		*Indians/Asians*		*Whites*		*National*	
Age group	*Census*	*DHS*	*Census*	*DHS*	*Census*	*DHS*	*Census*	*DHS*	*Census*	*DHS*
15-19	0.086	0.081	0.068	0.081	0.024	0.026	0.019	0.020	0.078	0.076
20-24	0.159	0.139	0.144	0.162	0.120	0.138	0.089	0.087	0.151	0.139
25-29	0.159	0.142	0.133	0.128	0.185	0.095	0.151	0.185	0.156	0.142
30-34	0.135	0.119	0.097	0.083	0.085	0.066	0.088	0.069	0.125	0.109
35-39	0.102	0.088	0.060	0.042	0.045	0.036	0.031	0.016	0.087	0.074
40-44	0.050	0.038	0.023	0.010	0.023	0	0.016	0	0.042	0.029
45-49	0.007	0.013	0.002	0.001	0.008	0	0.010	0	0.007	0.009
TFR	**3.49**	**3.11**	**2.64**	**2.53**	**2.45**	**1.80**	**2.02**	**1.88**	**3.23**	**2.89**

Source: Moultrie and Timæus (2003)

Note: The rates from the DHS were calculated using the reported numbers of births in the three years before the survey. Since the majority of the interviews in that survey were conducted in March and April 1998, the reference period for the resulting fertility rates is approximately October 1996.

Other authors, too, have presented their interpretations of fertility from these and other similar data. Udjo (2003), for example, estimates the level of African fertility to be 3.7 children per woman in 1996, for Coloureds 2.8, Indians/Asians 2.7 and Whites 1.9 children per woman respectively, and 3.2 children per woman nationally. The age distributions implied by his estimates of α and β, however, do not correspond at all well with the age distributions shown in Table 1.1, which are essentially the same for both sources of data. In particular, his estimates of older-age fertility are substantially higher. In an earlier version of the work based on the same data (and published as Udjo (1998)), he includes in the numerator all instances where up to six births in the previous year had been reported. Detailed interrogation of the 1996 census data (Moultrie and Timæus 2002) has shown that a large proportion of women in the 1996 census gave the same answer to questions on children ever born and births in the year before the census, in both cases reporting the number of children ever born, which – if not adequately controlled for – will inflate the fertility rates, particularly those of older women. Thus Udjo's approach is both biologically implausible and contains an implicit bias in the derived fertility rates, since the impact of the error increases with age. Furthermore, Udjo's estimates of the level of fertility from the 1996 data are suspect, especially with regard to his estimates of White and Indian/Asian fertility, since the standard schedule that underlies the Relational Gompertz model he applies to the data for these two groups is dissimilar to the age pattern of fertility demonstrated, and hence the results are likely to be distorted. The applicability of the Relational Gompertz Model is dealt with in greater depth in section 3.1.2.

The process of correcting the errors in the 1996 census fertility data has been described in other publications (Moultrie and Timæus 2002, 2003; Moultrie 2002), and is not repeated here. In any event, the errors observed in the 2001 census data are not the same as those found in 1996, and hence other strategies and corrections are called for. We have confidence in the estimates produced from the 1996 data. The age patterns of fertility demonstrated by the 1996 census data are highly consistent with those from the 1998 DHS despite the widely different methods applied to two very different sources of data. Hence, this report is premised on the assumption that the 1996 census data have been interrogated to such a degree that – unless strong and consistent evidence to suggest the contrary is forthcoming – assessments of subsequent fertility data must be compared with those earlier data and inconsistencies between those subsequent data and the 1996 data should be taken as being more probably indicative of errors in the more recent data than with 1996.

Section 2 examines the data collected in the 2001 South African census, and investigates the methods and mechanisms used to clean and edit the data, before concentrating on the fertility data collected. The edit rules and processes applied to the data emanating from the questions on fertility are subjected to particular scrutiny, with a view to determining the plausibility and usability of the fertility data collected. Particular errors relating to the data on women's reported parity (the number of children ever born) are identified and discussed. Having determined this, section 3 presents the derived estimates of fertility, disaggregated by population group, based on the 2001 census data. Section 4 extends the analysis to a provincial level for African South African women and presents estimates of the fertility of all women by province. By making use of these fertility estimates in conjunction with the reported parities and fertility estimates from the 1996 census, section 5 returns to the problem of misreported children ever born in Census 2001, and suggests and applies a method for correcting these data. The penultimate section compares the results derived here with those produced directly from the version of the census data to be publicly released with a view to determining the reliability of those direct estimates, and hence the reasonableness of the edits performed on the data. Section 7 summarises the results presented in the previous sections, and draws conclusions about the level of fertility shown by the data collected in the 2001 census. While this report is not the place to engage in a systematic and structured discussion on the public policy consequences of the results presented, specific results relating to teenage pregnancy and childbearing are discussed in greater depth. The report concludes with lessons on the usability and usefulness of the fertility data collected in the 2001 census and makes recommendations to Statistics South Africa regarding the collection, capture, editing, cleaning and release of fertility data in future censuses and surveys.

2 Data collection and cleaning in Census 2001

This section describes in detail the analysis and manipulations performed to produce robust and reliable estimates of fertility from the 2001 census data. To some readers, the detail may seem excessive and unnecessary. Nonetheless, this information is presented for three reasons. First, the descriptions set out here will permit other users to replicate the results derived. Second, the effects of the editing and data cleaning techniques used by Statistics South Africa (with assistance from the US Census Bureau) need to be carefully scrutinised in order to fully understand the effects and implications of their use on the data. Third, and related to the second point, the 2001 census was the first South African census to be subjected to such extensive data cleaning and editing procedures. For this reason, too, it is important to understand and evaluate the quality and effect of those procedures so as to inform decisions whether or not to use them in future censuses. Thus, detailed investigations into the data quality pursued here provide insights into whatever may have gone wrong with Census 2001, at least in respect of the fertility data, and hence may help avoid of similar errors in future censuses.

Section 2.1 introduces the editing and cleaning procedures applied to the 2001 census data, before examining the questions asked of women's fertility in section 2.2. Section 2.3 investigates the key stratificatory variables for examining the data on mothers (i.e. the denominator in the derivation of fertility rates). The data quality of sex, age and population group variables is interrogated. Section 2.4 discusses the variables used to derive the numerator used in the calculation of fertility rates, the numbers of births to women of a particular age in the preceding 12 months. Section 2.5 interrogates the data on children ever born and briefly discusses the data on the reported numbers of children still surviving.

2.1 The cleaning and editing process applied to the data collected in 2001

It is inevitable in any census that not all respondents will answer all questions fully or correctly. Further, not all responses (correct or otherwise) will be correctly recorded by the enumerator. Editing rules can be applied to the data to increase the amount of useful data available for analysis; to remove obviously inconsistent results (for example, a man reporting having given birth); for analytical tidiness (by removing the need to make a decision on what assumptions to make regarding missing data); and to ensure that missing cases are treated in a consistent fashion. By removing (or carefully editing) nonsensical responses, the overall quality and amount of 'useable' data may be increased.

The decision as to what to do in these circumstances is of critical importance in the census process. Three broad approaches can be identified. These can be used singly or in concert with each other. The first approach is to leave the data 'as is'. While this has the obvious appeal of minimising interference with the data, it is often not the best solution, since there will be situations where an answer to one question may be inferred logically from the respondent's (assumed to be correct) answers to other questions.

This approach ('logical imputation') relies on a series of 'rules', codified in a manual, to establish if other responses of the individual or those of other people in the same household allow missing or invalid responses to be inferred through logical reasoning. The validity of all such rules must be established and tested beforehand to avoid introducing biases into the data and because the extent of the intervention depends on the validity of the rules used.

However, such rules are unable to fill in all missing, incorrect or illogical responses. To do this, a further level of intervention in the data is required, known as 'hotdecking'. This method is less reliable than imputation, and thus should be used with caution. A hotdeck is used when the intention is to not have ANY missing entries in the data, regardless of whether an answer was given or could be determined on the basis of logical rules, in situations where a missing or invalid response cannot be corrected or inferred logically from the individual's responses to other questions, or from the response of other individuals in the same household. A hotdeck provides a mechanism to 'fill in the missing blanks' by attributing an answer to the respondent based on the response of the last respondent whose response was not subjected to logical imputation, and who has pre-specified attributes the same as those of the respondent for whom the response is missing. Clearly, unless one is extremely careful about the specification and population of such a hotdeck, biases can be introduced easily into the data.

The process adopted by Statistics South Africa in the cleaning and editing of the 2001 census data was slightly more complex than that outlined above. Statistics South Africa allowed a total of five possibilities for the way that each record of each variable could be treated in the editing process. First, the data could be left as it was recorded – i.e. the response was regarded as valid and (where possible) verified by the editing procedure. Second, the data could be subjected to logical imputation even though an answer had been given to the question recorded but where this answer was considered to be inconsistent with the answers to other questions. Thus, for example, if all answers but one to related questions are logically consistent, the inconsistent answer is changed to make it consistent with the others.

The third possibility allows for logical imputation to be applied where no answer is given. By again using 'plausible' answers to other questions, a logical 'answer' may be inferred. It stands to reason that the inherent danger exists in both the second and third approaches insofar as the acceptance of certain answers as 'valid' over others then determines the answers to subsequent questions.

The fourth and fifth approaches are based on the use of hotdecks. As with the logical imputation process, the drawing of a value from a hotdeck may be required where a response, inconsistent with all other related responses and not capable of being resolved logically, has been given to a question. Similarly, if no answer has been given and logical imputation is not possible, a hotdeck is required to complete the missing field. These two approaches are termed 'hotdeck from non-blank' and 'hotdeck from missing' respectively.

2.2 Fertility-related data collected in 2001

The census questionnaire in 2001 asked three questions of women aged between 12 and 50 from which it should be possible to assess levels and trends in fertility. The exact wording of the questions in the English questionnaire is shown in Figure 2.1. The first question (P-20) is designed to elicit the number of children ever born to a particular woman. This information, useful in its own right, is also important in the analysis of recent fertility data (given by the last question), especially where those data are considered (as is usually the case) to be defective. The last question (P-20b) should allow births that occurred in the twelve months before the census to be identified, regardless of the vital status of the child. The second question, on children still living, is not used directly in the measurement or analysis of patterns and trends in fertility but is important, together with data on children ever borne, for estimating infant and child mortality.

Figure 2.1 Wording of questions on fertility in Census 2001

P-20: TOTAL BIRTHS: How many children, if any, has (the person) ever had, that were born alive?
If none write 00 and go to P-21 (next question).
How many of these were boys?
How many of these were girls?
Include ALL her children, i.e. those who are still living, whether or not they live in this household, and those who are dead. DO NOT COUNT STILLBIRTHS (children born dead).

P-20a: STILL LIVING: If the person has ever given live birth:
If boys: How many boys are still alive?
If girls: How many girls are still alive?

P-20b: LAST CHILD BORN. If (the person) has ever given live birth:
When was (the person's) last child born? (DD/MM/YYYY)
What is the sex of that child? (M/F)
Is that child alive or dead? (A/D)
Write the day, month and year of the last live birth and dot the appropriate box of the sex. If multiple birth, indicate only the last child. Dot the appropriate box whether the child is still alive on Census night 9 - 10 October. DO NOT COUNT STILLBIRTHS (children born dead).

As will become evident, the questions on fertility were particularly badly answered by women in Census 2001. Even ignoring the questions on the sex composition of the numbers of children born and the numbers of children surviving, in aggregate more than half of all women of childbearing age had their answers to one or more of the following questions subjected to logical imputation or hotdecking: Total Children Ever Born; Total Children Surviving; day, month and year of last child's birth (three variables); last child's vital status and sex (two variables).

The data are summarised in Table 2.1 The high degree of imputation and hotdecking applied at younger ages is most probably due to the enumerators leaving the fertility responses of childless women (who tend to be younger) blank rather than writing zeroes (el-Badry 1961). Data on women under the age of 15 is seldom used in the estimation of fertility levels and trends, and so the extremely high levels of imputation and hotdecking in this age group are not cause for great concern. Even so, in no age or population group does the proportion of women whose data were not subjected to logical imputation or hotdecking exceed two-thirds. The apparent poor quality of the data among Indian/Asian and White women is more a reflection of comparatively greater levels of childlessness (leading in turn to a greater proportion of missing responses, especially at younger ages) than it is of poorer responses and enumeration of women in these two groups.

Table 2.1 Proportion of women whose fertility data were not subjected to logical imputation or hotdecking by age and population group, Census 2001

Age group	*Black African*	*Coloured*	*Indian or Asian*	*White*	*Total*
12-14	32.8	25.7	18.6	21.4	31.2
15-19	38.5	32.8	22.5	26.8	36.8
20-24	47.3	46.2	33.7	38.7	46.3
25-29	54.8	57.0	49.4	50.9	54.5
30-34	58.4	61.7	58.7	58.2	58.7
35-39	59.5	63.3	62.1	60.5	60.1
40-44	58.9	63.2	62.6	61.9	59.8
45-49	57.4	62.1	62.3	62.9	58.8
Total	**49.6**	**51.0**	**46.5**	**49.3**	**49.6**

Interestingly, too, the quality of the data varies little by the level of mother's education[1], thereby eliminating one possible source of heterogeneity in the evaluation of the quality of the fertility data in the census (Table 2.2). While there is a marginal improvement in the proportion of women with unedited fertility data with increased levels of education (in part due to the confounding effects of younger women having less, or uncompleted, education and younger women having generally poorer fertility data on account of a larger proportion of them being childless), this increase is small. Only among the very few women with higher education (4.7 per cent of Africans, 4.2 per cent of Coloureds, 13.5 per cent of Indian/Asians and 26.3 per cent of Whites) is the proportion of unedited fertility data noticeably higher.

Table 2.2 Proportion of women whose fertility data were not subjected to logical imputation or hotdecking by education level and population group, Census 2001

Education level	*Black African*	*Coloured*	*Indian or Asian*	*White*	*Total*
No schooling	50.5	51.0	43.4	36.1	50.4
Some primary	46.0	48.0	38.3	24.1	45.7
Primary	47.7	50.1	39.5	25.6	47.2
Some secondary	49.9	52.2	47.3	43.8	49.7
Matric/Std. 10	51.9	50.8	47.5	53.5	51.9
Higher	57.1	56.2	48.6	56.1	56.3
Total	**49.6**	**51.0**	**46.5**	**49.3**	**49.6**

Across all population groups, the general trend is for the extent of imputation necessary to decrease with both age and level of education, although the data presented in Table 2.3 suggests that the quality of data collected among women aged less than 30 with a matric or a post-secondary education is noticeably worse than those of other women with less education.

The extent of the imputation and hotdecking of the fertility data (even at older ages and among the most educated women) is cause for concern, and points to a worrying lack of training on the part of enumerators. (Only about 4 per cent of questionnaires were completed by respondents and not by enumerators.) Given the consistency of the magnitude of the errors across all social and educational strata, the most plausible explanation is that inadequate training or the design of the 'cascade'-type enumerator training process employed is responsible. This is a matter that requires urgent attention and redress before the next census is conducted in South

[1] Incidentally, there is a flaw in the (derived) "education groups" variable prepared by Statistics South Africa. People with a matric (Grade 12) are counted twice, once in the group "Standard 10/Grade 12", and a second time in the group "Higher". This needs to be addressed as a matter of urgency.

Africa. In this regard, particular attention should be directed at the training manuals used to instruct enumerators.

Table 2.3 Proportion of women whose fertility data were not subjected to logical imputation or hotdecking by education level and age group, Census 2001

Age group	*No Schooling*	*Some primary*	*Primary*	*Some secondary*	*Matric*	*Higher*	***ALL***
12-14	26.5	31.3	31.2	31.1	0.0	0.0	**31.2**
15-19	38.0	36.8	38.1	36.8	35.5	35.4	**36.8**
20-24	42.9	46.2	50.1	48.8	44.3	42.4	**46.3**
25-29	48.5	54.0	56.6	57.1	53.8	52.6	**54.5**
30-34	52.3	57.5	60.1	60.7	58.9	60.0	**58.7**
35-39	53.9	59.6	61.9	62.0	60.9	62.5	**60.1**
40-44	53.8	59.4	61.8	62.0	61.3	63.4	**59.8**
45-49	53.2	58.4	60.8	61.0	61.4	64.1	**58.8**
ALL	**50.4**	**45.7**	**47.2**	**49.7**	**51.9**	**56.3**	**49.6**

Statistics South Africa was aware of the problems before the census went into the field: the Pilot Census had shown a non-response rate of over 23 per cent, and required imputation rates of between 20 and 25 per cent (Statistics South Africa 2003: 14). The extent of the imputation of the fertility data is also worrying for other reasons. Statistics South Africa initially set a target cut off for imputation of two per cent:

> Additionally, it is important to ensure that the imputation rates for any particular variable are never high enough to introduce bias in the data. The final data set will carry a set of flags that will indicate whether or not each variable was modified by the editing system. This will permit a thorough analysis of the effects that editing has had on the data. Any edit that causes a variable to be imputed more than, say, 2% of the time should be investigated and reviewed with analysis. (Statistics South Africa 2003: 13)

Interestingly, however, the final version of the edit specifications (Statistics South Africa 2003) omits this paragraph entirely. In addition, the reported levels of imputation are inexplicably somewhat lower than indicated in Table 2.1, although still unacceptably high. It is claimed that the overall non-response rate on the "fertility section" (an undefined term) was 13.5 per cent among women aged 12 to 50, with logical imputation being applied in between 1 and 21 (!) per cent of cases, and hotdecking approximately 11 per cent of the time (Statistics South Africa 2003: 16). In addition to the figures cited above, the final edit specifications simply observe that the non-response rate declines with age of woman, adding

> Caveat: While it is true that proper editing will reduce content error, it is also true that *improper* editing can introduce further response error. Statistically speaking, imputation rates must be kept low in order to avoid biasing the data." (Statistics South Africa 2003: 10, emphasis in original)

In the case of the fertility data, imputation rates are not low. Given the extent of imputation and hotdecking employed, it is necessary to pay particular attention to the effects of the edit rules used to impute and edit women's fertility data, and to ensure that the specification and population of the hotdecks employed do not introduce biases into the data. The edit procedures used by Statistics South Africa under the guidance and advice of the US Census Bureau do not, unfortunately, inspire much confidence. The logical imputation rules make extensive use of the "mother person number" (MPN – a reference to the line number of the questionnaire that contains the data on a respondent's mother). This is worrying since, by Statistics South Africa's own admission, the MPN together with other 'linking' variables is problematic and offered "a major challenge to the editing team" (Statistics South Africa 2003: 16). The MPN of approximately one in every six infants was imputed, this proportion being relatively constant by population group. Among all children under the age of five, the MPN was imputed in between 11.1 per cent of cases (Indians/Asians) and 15.9 per cent of cases (Africans). *Edits based on MPN, then, should be treated with a degree of scepticism* and – unfortunately – the fertility edits are heavily dependent on this variable.

The construction and population of the hotdecks is of even greater concern. A case in point is that of determining a woman's parity (children ever born) when no answer is given. As suggested above, it is a well-established fact that women's parity is differentially underreported by age. This happens because childless women (who also tend to be younger) are frequently enumerated as 'parity missing' rather than 'parity zero' in censuses and surveys, an error brought about by the enumerator failing to write zero in the answer to the parity question, and leaving the relevant box blank.

The edit rule constructed for use with the Census 2001 data shows no sensitivity to this fact. A very large proportion of younger women's parities was missing in the raw data. Where evidence of parity could not be inferred logically (the majority of cases), the woman's parity was imputed from a hotdeck. However, it stands to reason (for exactly the reasons outlined above) that the contents of the hotdeck from which parity would be imputed will be biased towards women who have had a child. As a result, the "corrected" parities for the youngest women in the 2001 census are higher than they should be, since the hotdeck has the effect of imputing children to evidently childless women. This is discussed further in section 2.5.

A further peculiar characteristic of the hotdeck applied to resolve women's parity is that in determining parities for never married women, only those never married women with three or fewer children are used to establish the hotdeck. This constitutes a systematic bias against older never married women with more than three children. The proportion of women who have never married in South Africa is high: according to the data collected in the 1998 South Africa DHS, almost one in three African South African women aged between 30 and 35 had yet to marry, compared to approximately one in every twenty White South African women who had never married by that age. Even among the oldest women (those aged 45 to 50), nearly one in every seven African women had never married, while almost all White women had been married at least once.

The implicit assumption that all never married women have had three or fewer children is unreasonable, even though in aggregate it may appear otherwise. According to the 2001 census data, 96.2 per cent of all never married women aged 12-49 had had three or fewer children, ranging from 95.8 per cent among Africans to 99.8 per cent among Indians/Asians and Whites.

As can be seen from Table 2.4, more than three in every ten never married African women aged between 45 and 49 have had more than three children. Even among African women aged between 35 and 39, almost a fifth has had more than three children.

Table 2.4 Proportion of never married women with three or fewer children by age and population group, Census 2001

Age group	*Africans*	*Coloureds*	*Indians/Asians*	*Whites*	*TOTAL*
12-14	100.00	100.00	100.00	100.00	100.00
15-19	99.95	99.98	99.98	99.99	99.96
20-24	99.62	99.85	99.93	99.93	99.66
25-29	97.62	98.81	99.78	99.83	97.82
30-34	90.59	95.46	99.17	99.50	91.38
35-39	80.13	89.90	98.25	99.10	81.64
40-44	72.84	86.93	97.91	98.18	75.07
45-49	69.15	84.29	97.53	97.65	71.74
Total	**95.8**	**98.0**	**99.8**	**99.8**	**96.2**

Thus, in some instances at least, it would appear that the designers of the editing rules paid little or no attention to the underlying demography of the South African population. The errors introduced into the data are non-trivial, and should be corrected before the data are made available to the public.

2.3 Demographic data on mothers

The material presented above suggests that even the choice of records for inclusion in the assessment and analysis of the fertility data is not without contention. Since each and every demographic variable, including age, sex and province of residence was potentially open to logical imputation and hotdecking, it *behoves the careful demographer to investigate the extent to which these central demographic variables were, indeed, subjected to cleaning and editing, before using them as stratificatory variables in subsequent analyses.* In this section, the data on population group, sex and age are discussed. Details of the extent to which each were subjected to imputation and hotdecking are shown.

In all cases, unless explicitly stated to the contrary, the data presented in this section apply the person adjustment factors ('weights') to the data to correct for the estimated undercount of the South African population in Census 2001.

2.3.1 Population group

Demographic outcomes vary in crucial ways in South Africa by population group. In essence, these outcomes reflect the harsh realities, and long-term institutional consequences, of apartheid and apartheid policies. Thus, even though there is very little evidence of racial determinants of demographic variables (such as fertility and mortality) beyond the effect of other social variables, it is imperative to analyse demographic data by population group.

Table 2.5 shows that all those who gave their population group as African, Coloured, Indian/Asian or White were not subjected to data cleaning or editing in respect of their population group. Those who gave their population group as 'Other' were more likely to be allocated a population group via a hotdeck (68.6 per cent of 94 000 people), while those who did not answer the question ('undefined') were allocated to a population group primarily through logical imputation based on the population group of the head of the household, the population group of other household members and language spoken in the household. Where these variables could not be used to impute a population group, a hotdeck was used (Statistics South Africa 2003: 156).

Table 2.5 Edited and unedited population group and imputation flags in Census 2001, all South Africans

Original classification and edit procedure used		African	Coloured	Indian/Asian	White
		Final Classification			
Unchanged		35,023,912	3,906,853	1,096,893	4,216,633
Other					
	Logical - blank	-	-	-	-
	Logical - non-blank	19,181	4,153	3,793	2,345
	Hotdeck - blank	-	-	-	-
	Hotdeck - non-blank	13,831	29,224	3,998	17,366
	Total	33,012	33,377	7,791	19,711
Undefined					
	Logical - blank	321,066	26,019	6,749	29,362
	Logical - non-blank	4,479	407	85	300
	Hotdeck - blank	33,510	27,539	3,904	27,329
	Hotdeck - non-blank	187	311	46	305
	Total	359,242	54,276	10,784	57,296

Several problems are apparent in the data on population group. First, respondents were offered a response of 'Other, please specify' as an alternative to the four major population groups. These 'other' responses were not coded separately (e.g. Griqua as distinct from Khoi) even at the raw data stage, and – as can be seen from the table – all these 'others' were allocated to one of the four principal population groups in the final coding of the data, with a preference for Coloured. While the numbers are statistically unimportant, it is important that if a census is to generate feelings of social inclusivity and be free of suspicion, then questions that offer a possible response of 'other, please specify' should not be included if these responses are going to be replaced with others determined by editing rules. This is a matter that deserves careful attention in the planning of future censuses.

Having said this, it would appear that, in aggregate, the data by population group have not been subjected to a significant degree of imputation or hotdecking, although the extent of imputation or editing may have been more severe at particular combinations of age and sex or in particular provinces. However, with only 1.3 per cent of the South African population's population group subjected to editing or cleaning, the effects of these changes will be trivial. *In all subsequent analyses, then, the cleaned and edited population group variable (P06) is used.*

2.3.2 Sex

We investigated the extent to which a woman's sex[2] was determined by the editing rules, by (cleaned) population group. The editing rules to determine sex are complex, and involve separate rules for the determination of the sex of the head of the household relative to other household members. In addition, some of the final resolutions of sex are made contingent on variables that are known to be badly answered (the MPN and FPN, particularly).

While a number of individuals who reported themselves as being of one sex were edited to be of the other according to this algorithm, which should be carefully revisited before it is used again, the extent of the imputation or hotdecking is extremely small, for all population groups. Less than 0.2 per cent of recorded men in any population group were recoded as women during

[2] Sex refers to the biological attributes of a person, classified as male or female. Gender, on the other hand, refers to a sociological, and frequently self-classified, construction based on sexual identity. Hence, in the discussion of demography, it is more correct to refer to 'sex' rather than 'gender'.

the editing stage, while less than 0.06 per cent of women recorded as female were recoded as male. The proportion of respondents for whom sex was blank was small, ranging from 1.47 per cent among Coloureds to 1.00 per cent among Africans. These 'undefined' cases were subject to logical imputation 75 to 80 per cent of the time, the balance being attributed a sex by means of a hotdeck. In total, though, the edits applied to the variable are inconsequential. *Accordingly, the cleaned and edited data on sex were accepted for use in the analysis, despite the reservations expressed above.*

2.3.3 Age

The census questionnaire asked respondents both their date of birth, and their age at their last birthday. Investigation of the data indicates that the latter was subject to a far greater degree of error, and hence imputation and cleaning, than date of birth. Each of the three components of the date of birth (i.e. day, month and year of birth) was assessed separately for the extent of imputation among women by population group.

Some answers to the questions were patently nonsensical (e.g. date of birth reported as 31 February). The edit manual somewhat surprisingly made an explicit choice to ignore such dates, by instructing to not "correct for day/month inconsistencies such as February 30, April 31, etc." (Statistics South Africa 2003:105). Although not a big problem on its own, the error was compounded in that a record with such an answer was then deemed to be 'reasonable' and hence added to the hotdeck pool with the result that a number of records were given nonsensical birth dates as a result of using the flawed hotdeck. *All records pertaining to women of childbearing age with birthdates that are not possible in the Gregorian calendar (approximately 24 000 cases) were excluded from the data used to estimate fertility, although this number is unlikely to affect the estimates of fertility derived to any great extent.*

The data are not easily tabulated. However, summary statistics on the extent of imputation and hotdecking of data on the day, month and year of birth of women adjudged (in the final release of the census data) to be aged between 12 and 49 are shown in Table 2.6.

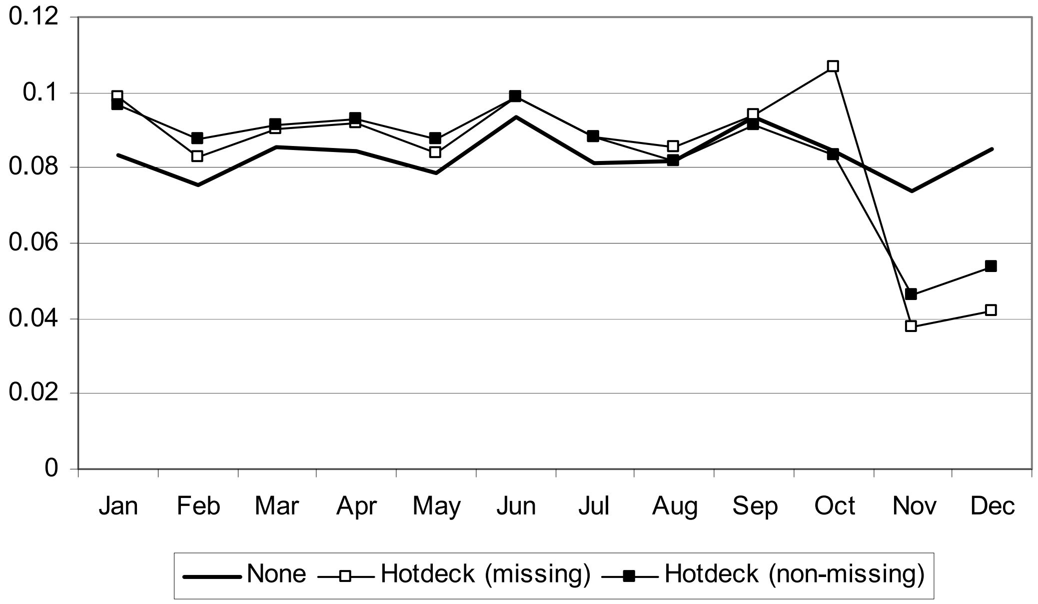
Table 2.6 Extent of editing and cleaning of women's date of birth (percent of cases), women aged 12 to 49

	No imputation	*Logical imputation applied to*		*Hotdeck applied to*	
		missing response	*non-missing response*	*missing response*	*non-missing response*
Day of birth	95.8	n/a	n/a	4.2	< 0.1
Month of birth	95.8	n/a	n/a	4.0	0.2
Year of birth	97.7	1.0	1.3	< 0.01	< 0.001

Note: n/a : not applied

Slightly more than 4 per cent of the 14.1 million women reported as being aged between 12 and 49 inclusive reported an invalid day of birth and were recoded using a hotdeck (there being no means of attributing an answer on logical grounds). Similarly, month of birth was imputed in 4.2 per cent of cases (although obviously not entirely the same women for whom day of birth was imputed). Only 2.3 per cent of these women were accorded a year of birth different from that which they had reported to the enumerator. Of course, reporting a logically valid day, month or year of birth does not avert heaping on certain digits, and this much is readily apparent in the data. The distribution of reported days of birth is not uniform: there is a clear preference for days early in the month, with an obvious preference for the first of the month, as well as the 10th (the census day). The distribution by month of birth was remarkably consistent among the data not subjected to hotdecking. However, there is an obvious (and potentially serious) error in the hotdeck construction that reduces the likelihood of being allocated a birth month of either November or December. The source of this error is unknown, and needs to be investigated[3].

Figure 2.2 Distribution of women's reported month of birth by logical imputation and hotdeck status, Census 2001



Finally, in terms of the reported year of birth, there is an apparent problem with the logical rules that allocate a year of birth from 'non-missing' (but implausible) data, particularly at either extreme of the age range, as the logical rules inadvertently contain a bias against the allocation of a birth year after 1986. An extremely small proportion of birth years were subjected to hotdecking. This comes as a relief, as the hotdecks are so biased as to be useless: of the nearly 600 women aged 12 to 49 whose year of birth was subjected to hotdecking, a quarter (151) were allocated 1984 as the year of birth, the vast majority of whom were subsequently identified as White. This is a clear example of the dangers of using a hotdeck.

The conclusion reached, therefore, is that – despite some identified errors and issues with the hotdecking and imputation rules – the edited and cleaned data in respect of these three demographic variables as they apply to women of reproductive age can, and should, be used in other tabulations.

2.4 Data on children

While the basic demographic data on mothers are generally good, the same cannot be said of the data on their fertility. As indicated earlier, one of the three questions on fertility asked the date of birth of the last child born. The equivalent question in Census 1996 asked how many births had occurred in the twelve months before the census. That question was poorly answered, and

[3] This is a different situation to that relating to children, since children known (or imputed) to be born in 2001 could not have been allocated a birth month after the census date. In the case of women, no such restriction applies to their birth month.

Census 2001 tried to address this by asking simply the date of the last child's birth. The resulting data, by day, month and year of birth are examined and discussed below.

2.4.1 Day of last child's birth

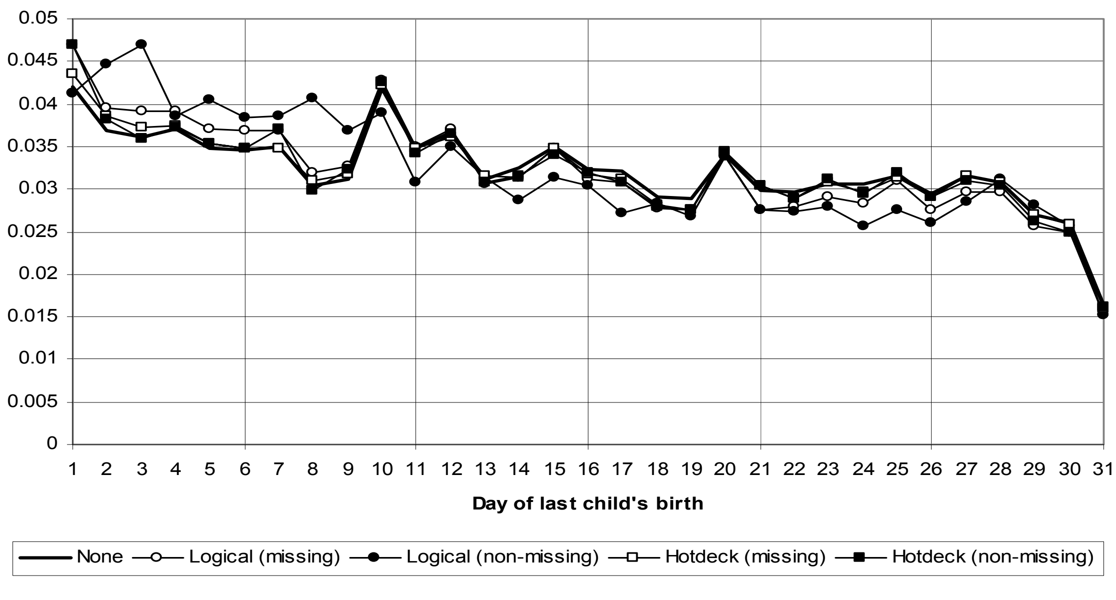
Table 2.7 shows the number and proportions of women aged between 12 and 49 by the edit status of their response to the day of their last child's birth.

Table 2.7 Distribution of women aged 12 to 49 by imputation flag for response to question on day of last child's birth

	No	*Logical imputation applied to*		*Hotdeck applied to*		
	Imputation	*missing response*	*non-missing response*	*missing response*	*non-missing response*	*TOTAL*
Women	6220655	751833	584423	827046	71772	8455729
(per cent)	73.6	8.9	6.9	9.8	0.8	100

More than one in four women of reproductive age (26.4 per cent) did not have a plausible response recorded to this question. Blank responses that had to have a day of birth imputed or hotdecked accounted for 8.9 and 9.8 per cent of all responses respectively. A further 7 per cent of women gave nonsensical responses, and had to have their responses imputed based on the responses for the household member 'identified' as being that woman's child. Among those women whose last child's day of birth was not subjected to imputation or hotdecking, the distribution of days of birth was almost uniform, although there was an indication of a clear preference for days early in the month (especially the first), and for the census day (Figure 2.3).

Figure 2.3 Distribution of last child born's day of birth by imputation and cleaning method, Census 2001



The reasonableness of the imputation and hotdecking procedures applied to these data can be assessed by comparing the distributions of days of birth by imputation method with the distribution of the unedited and uncleaned data. For this variable, the hotdeck procedures evidently worked well. However, there are clear and obvious biases in the distribution of birth days for data subjected to logical imputation based on non-missing data. The logical imputation procedure for missing data also favoured dates before the census night, which is difficult to explain. A fundamentally similar pattern is observed when only the birth days of those births that occurred in the year before the census are scrutinised.

2.4.2 Month of last child's birth

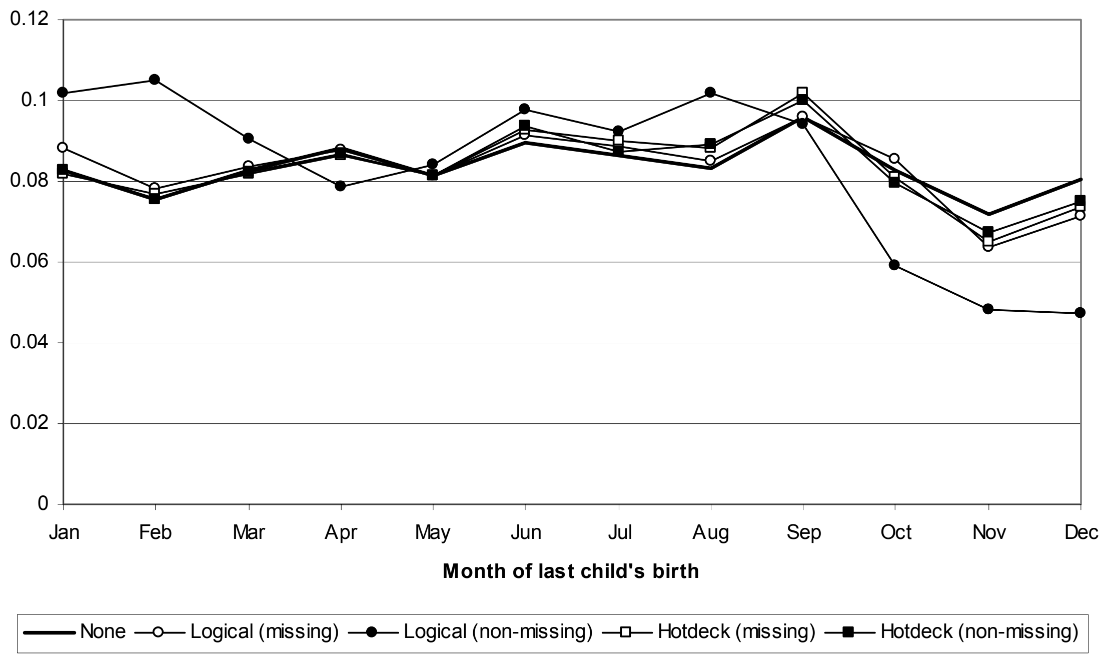
Table 2.8 shows the number and proportions of women aged between 12 and 49 by the edit status of the response to the day of their last child's birth.

Table 2.8 Distribution of women aged 12 to 49 by imputation flag for response to question on month of last child's birth

	No imputation	*Logical imputation applied to missing response*	*Logical imputation applied to non-missing response*	*Hotdeck applied to missing response*	*Hotdeck applied to non-missing response*	*TOTAL*
Women	6230403	748555	577534	822614	76622	8455728
(per cent)	73.7	8.9	6.8	9.7	0.9	100

Again, more than a quarter of women aged 12 to 49 did not have a plausible response recorded to this question. In general, the imputation and hotdeck procedures worked well (Figure 2.4), except in the case of logical imputation where there was a non-missing response (6.8 per cent of cases). In these circumstances, the rules applied contain an evident bias against birth months in October, November or December, and bias towards January and February.

Figure 2.4 Distribution of last child born's month of birth by imputation and cleaning method, Census 2001



This bias arises, in part, due to the need to impute a year of birth to an infant in many cases, and if that year was set to the census year (2001), then it was not possible for the month of birth to be set to November or December, although this cannot explain why the same phenomenon was not also observed in edits arising from the application of a logical rule applied to missing data.

The error with the logical (non-missing) data points to a substantive source of error in the data: that since each variable is coded and cleaned without reference to the editing processes to which other variables have been subjected, once a value has been set (either through a valid response, or cleaning and editing), that value is then used in the determination of values for other variables. This is something that will have to be investigated more carefully in the specification of editing rules in future censuses.

2.4.3 Year of last child's birth

The reported year of the last child's birth is of fundamental importance in estimating the total numbers of births that occurred in the twelve months before the census, and hence recent fertility rates. Table 2.9 shows the number and proportions of women aged between 12 and 49 by the edit status of their response to the year of their last child's birth.

Table 2.9 Distribution of women aged 12 to 49 by imputation flag for response to question on year of last child's birth

	No	*Logical imputation from*		*Hotdeck applied to*		
	imputation	*missing response*	*non-missing response*	*missing response*	*non-missing response*	*TOTAL*
Women	6560661	604260	391548	734257	165002	8455728
(per cent)	77.6	7.1	4.6	8.7	2.0	77.6

Several issues are apparent. First, it would be wrong to infer, on the basis that the total number of births is significantly below one million, any extent of an under-recording of infants from these data, as the data for 2001 refer to only nine and a half months. However, of far greater concern is the fact that, of those reported in the edited data to have been born in 2001, almost three out of every ten were imputed (Table 2.10).

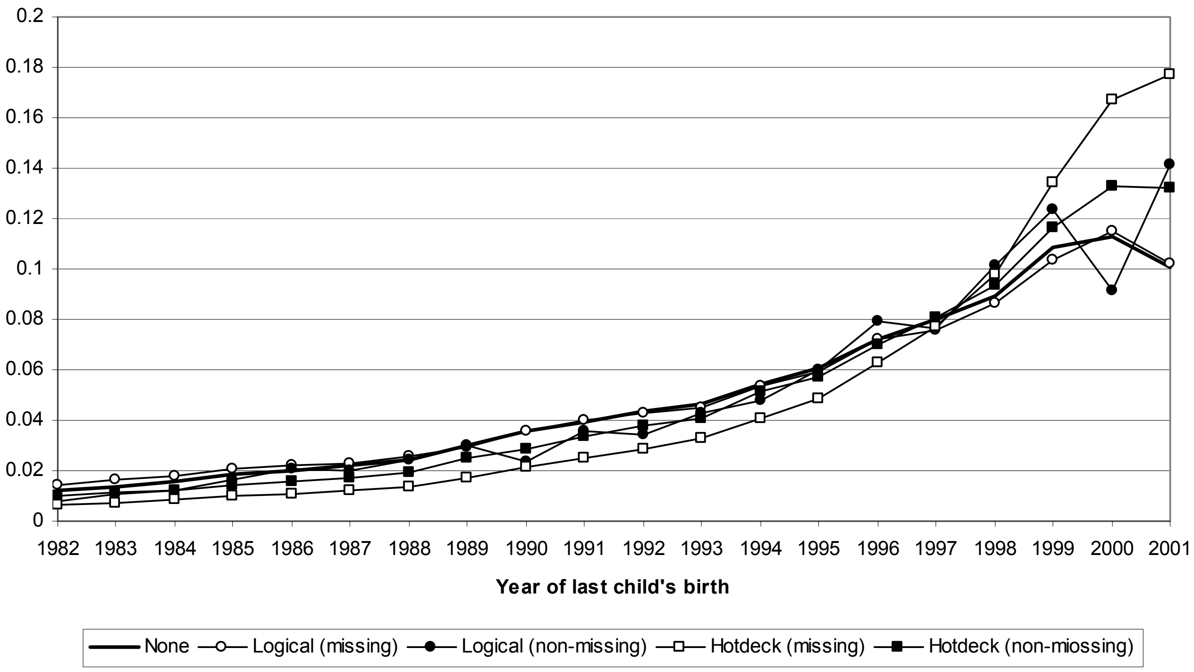
Table 2.10 Distribution of women aged 12 to 49 by imputation flag for response to question on year of last child's birth, births reported as occurring in 2000 and 2001

	No	*Logical imputation from*		*Hotdeck applied to*		
Year	*imputation*	*missing response*	*non-missing response*	*missing response*	*non-missing response*	*TOTAL*
2000	74.6	6.9	3.7	12.6	2.2	947597
2001	70.8	6.5	6.0	14.2	2.4	889350

Moreover, a propos this variable, the edit and hotdecking procedures would appear to be seriously flawed. Figure 2.5 shows the distribution of years of last child's birth in the twenty years before the census by logical imputation and hotdecking status. The vast majority of last births to women of childbearing age occurred after 1982, and hence those occurring before 1982 have been excluded from the figure (but not the analysis) simply for ease of representation. In conjunction with the low proportion of the data not subjected to editing or imputation presented in the table above, it would appear that the editing and imputation rules play a significant role in determining the numbers and patterns of births in the years before the census.

First, it is immediately apparent that the "hotdeck applied to missing" routine attributes a much larger number of births than expected to the years immediately preceding the census, and fewer to earlier years. Second, the "logical from non-missing" algorithm evidently contains an odd bias against births in 2000. Of still greater concern is that the extent of imputation is highest in the years immediately preceding the census (Table 2.10).

Figure 2.5 Distribution of last child born's year of birth (last births in the twenty years before the 2001 census) by imputation and cleaning method, Census 2001



Both the extent of and the bias in the imputation of last born children's year of birth make it incumbent on any demographer or other user of these data to pay extremely careful attention to issues of data quality, and decisions of which edits to accept, and which to reject. The process of deciding which fertility data to admit in the analysis of recent trends in South African fertility is discussed in greater detail in section 3.

2.5 Data on women's reported parity

The data on children ever born are more fraught with error than those relating to their mothers. Table 2.11 shows the proportion of women, by age and population group, whose parity data were not subjected to data editing and cleaning.

Table 2.11 Proportion of women whose parity data was not subject to logical imputation or hotdecking, by age and population group, Census 2001

Age group	*African*	*Coloured*	*Indian/Asian*	*White*
12-14	65.2	53.5	61.4	46.2
15-19	73.5	63.7	68.8	55.9
20-24	82.5	78.5	79.1	73.9
25-29	88.2	87.6	88.0	85.4
30-34	90.9	91.2	92.2	90.2
35-39	91.9	92.6	93.5	91.3
40-44	91.4	92.5	93.3	91.5
45-49	89.9	91.3	91.9	90.4

While the trend by age is to be expected, as younger women are more likely to be childless, and hence may not complete the fertility section fully, the magnitude of the adjustments applied has to make one extremely cautious in using these data. To add to the concerns, the vast majority of the cleaned and edited cases had to be hotdecked from blank (in no population group or age group did the proportion of data cleaned by the other three methods together exceed 4 per cent of the total). Interestingly, too, the quality of the data collected worsens with age after 40 among Africans, Coloureds and Indians/Asians.

The results presented in this table offer a more accurate interpretation of the same data to that offered by Phillips, Cronje and Phoshoko when they observe that the "average percentage of CEB 'not stated' (in) 2001 was approximately 10 percent" (2003:2). While this is indeed the case for the average, the parity data are of exceedingly poor quality at younger ages, with parity data in respect of one in every five women aged under 25 having to be deduced or imputed.

This edit should not have been implemented. In 1961, el-Badry observed that the relationship (by age) of the proportion of childless women and the proportion of women whose parity is not stated is approximately linear. Based on this relationship, he derived a method of correction that could be applied to the data which would estimate the "true" proportions childless, and the "true" proportions of unstated parity (el-Badry 1961). This technique, named after its originator, is a recognised and valuable tool available to demographers, and features in the standard reference work of indirect techniques (United Nations 1983). The downside of the approach, of course, is that it can only be applied to aggregate data, and hence cannot be used as a means of 'filling in the missing blanks' in the individual level data.

Apart from this, there is an inherent danger in using a hotdeck with parity data, namely that the characteristics of the women used to populate the hotdeck from which a value is chosen may be materially different from the group of women to whom the hotdeck is applied. This is particularly the case at younger ages. Women who have had a child are more likely to have completed some or all of the fertility section, thus increasing the likelihood that they are included in the hotdeck. Circumstantial evidence of this effect possibly being at play in the 2001 census data comes from the fact that while the average parities of women whose parities were drawn from the hotdeck are generally consistently and markedly lower than those of women whose parity data were not subjected to cleaning and imputation, the reverse holds at the two youngest age groups. Thus, for example, among African women aged 12-14, the average parity among hotdecked women is more than twice that among women whose parity data were not subjected to cleaning or imputation. Among teenage women 15-19, the average parity of women subjected to hotdecking was 12 per cent higher. In both cases, the reverse pattern would have been expected had the hotdeck been working properly. A similar pattern was observed in the data for young Coloured women.

However, there is a far more serious problem with the 2001 census data on children ever born: the data from 1996 and 2001 are logically incompatible. This is easily shown as follows. The two censuses were held exactly five years apart. Hence, women who were aged between 45 and 49 in 2001 would have been aged 40-44 in 1996. The average parity of women in a given age group is derived by dividing the total number of births to women in that age group by the number of women. Because women cannot "unhave" children, it follows that a cohort's average parity must increase as it ages, the exception to this being recent significantly higher mortality of higher parity women relative to lower parity women, which is implausible, and for which no evidence can be marshalled.

Table 2.12Table 2.12**Error! Reference source not found.** presents data from the two most recent South African censuses.

Table 2.12 Unadjusted ('raw') and adjusted average parities in 1996 and 2001, by population group

	Africans				
	1996		*2001*		
Age group	*Unadjusted*	*Post-el-Badry*	*Unadjusted*	*Post-Edit*	*Post-el-Badry*
12-14	0.04	0.02	0.03	0.01	0.02
15-19	0.23	0.16	0.20	0.19	0.16
20-24	0.91	0.75	0.77	0.74	0.67
25-29	1.74	1.58	1.48	1.45	1.38
30-34	2.69	2.55	2.31	2.28	2.23
35-39	3.48	3.35	3.07	3.05	3.01
40-44	4.16	4.05	3.59	3.57	3.51
45-49	4.61	4.50	3.94	3.92	3.81
			Coloureds		
12-14	0.03	0.02	0.03	0.01	0.01
15-19	0.18	0.14	0.20	0.20	0.13
20-24	0.79	0.68	0.76	0.74	0.62
25-29	1.52	1.42	1.43	1.41	1.30
30-34	2.24	2.16	2.06	2.03	1.94
35-39	2.82	2.75	2.53	2.51	2.42
40-44	3.27	3.19	2.84	2.81	2.72
45-49	3.74	3.66	3.06	3.04	2.90
			Indians/Asians		
12-14	0.03	0.02	0.02	0.00	0.01
15-19	0.07	0.05	0.07	0.05	0.05
20-24	0.53	0.42	0.39	0.34	0.32
25-29	1.34	1.20	1.08	1.04	0.99
30-34	2.06	1.94	1.80	1.77	1.73
35-39	2.43	2.32	2.25	2.23	2.20
40-44	2.71	2.61	2.46	2.43	2.40
45-49	2.92	2.78	2.57	2.55	2.48
			Whites		
12-14	0.02	0.01	0.03	0.01	0.01
15-19	0.06	0.04	0.06	0.05	0.04
20-24	0.35	0.29	0.31	0.26	0.24
25-29	1.02	0.92	0.88	0.84	0.79
30-34	1.67	1.58	1.51	1.48	1.42
35-39	2.05	1.96	1.92	1.90	1.84
40-44	2.24	2.15	2.11	2.09	2.03
45-49	2.40	2.31	2.20	2.19	2.10

To better ensure comparability, the data from each census have been prepared in identical ways. The data from 1996 are presented in two variants, the first being the actually-recorded parity of each woman. While no detailed description of the editing rules applied in 1996 has been published, it would appear that the data on parity were not subjected to strenuous editing, since several instances of very young women reporting implausibly high (five or more) numbers of births can be found. A significant proportion of women's parities were not recorded, and hence an el-Badry correction was applied to the data (Column 2).[4]

[4] In the final analysis of the 1996 census data, further allowance was made for the erroneous inclusion of stillbirths among a woman's live births, further reducing average parities. The adjusted parities after application of this correction are not presented here, although use is made of them in section 5 to derive corrected parities from the 2001 census data.

Three variants of the 2001 census data are presented. The first is equivalent to the unadjusted 1996 data. The second shows the average parity of women after editing and imputation – the edited data as prepared by Statistics South Africa. The final column shows the estimated average parities after application of an el-Badry correction to the unadjusted data. (Note that it is incorrect, and in fact meaningless, to apply an el-Badry correction to the edited data since, first, there are no longer any missing data, and second, the editing rules make their own assumptions about the fertility of women whose parity data are missing).

Several aspects relating to the parity data in the 2001 census are apparent in Table 2.12. First, the edit rules applied to the 2001 census data do not reasonably emulate those produced using the el-Badry correction, suggesting that the edit rules do not fully capture the known nature of the relationship that exists between childlessness and missing parity data. This issue is revisited again in section 3.2.1.

Second, the unadjusted parities reported by women aged 45-49 in all population groups in 2001 are lower than the equivalent parities reported by women aged 40-44 in 1996. This is indicative of a clear inconsistency in the parity data between the two censuses which, although not as obvious, extends to the lower ages. As a result, all demographic techniques that use reported lifetime fertility as an input (Brass P/F methods and its associated variants, Relational Gompertz models, Brass Children Born-Children Surviving methods) need to take these contradictions into explicit account.

The implausible but nonetheless possible explanation that the parity data collected in the 1996 census are overstated could be advanced to explain the discrepancy. Udjo (2003) suggests precisely this in relation to some population groups and provinces: "the 'best' estimates based on the 1996 census still overestimate fertility in most provinces because of overreporting of children ever born". However, this quote references another of his publications (Udjo 2003:421), in which he comes to rather different conclusions, namely

1. That the 'curving upward' of the reported average parities of older African women is indicative that women may have <u>under</u>reported their fertility in the <u>1995 October Household Survey</u> (emphasis added);
2. That there may be "some degree" of overstatement of parity among non-African women in the <u>1995 October Household Survey</u> (emphasis added).
3. That the 'curving downward' of the P-points in the census should be attributable to age-exaggeration.

However, age exaggeration would imply an understatement of average parities, not an overstatement, as the average parities in the affected age groups would be reduced by the lower average parities of younger women included as a result of age exaggeration.

In addition to the obvious contradictions, it is hard to see how White, Coloured, and Indian/Asian women may have over-reported their lifetime fertility in 1996. The more reasonable assumption alluded to by Udjo is that these women may have understated their ages (a likely sociological phenomena in these groups), resulting in higher parities for a given reported age than one may have expected. Until some corroborating evidence is produced, assertions that the discrepancies can be explained by assuming that the 1996 data are overstated should be treated with scepticism. In any event, conventional demographic wisdom points to the fact that women in developing countries habitually underreport their lifetime fertility in censuses and surveys (Brass 1996; Cleland 1996).

Four other possibilities exist to explain the incompatibility, the first three relating to the relative quality of the data in the 2001 and 1996 censuses. The first three are that the parity data from both censuses may be wrong, or the parity data from one of the censuses may be correct while that from the other is wrong. Two of these three explanations can probably be rejected. The 1996 census parity data have been subjected to rigorous analysis and investigation, and the results have been largely confirmed by the results from the 1998 South Africa DHS (Moultrie and Timæus 2002). Further, as shown later, the parity data from the 1996 census are broadly consistent with the levels of fertility estimated. Thus, unless other evidence is found to assert that the parity data in 1996 were seriously at fault, we must conclude that it is the more recent parity data that are wrong.

The final explanation is that there has been significantly and generalised differential mortality of higher parity women relative to lower parity women in the last five years, thereby reducing average parities. This possibility is unlikely given the uniformity of findings across population groups. *The overwhelming evidence, then, points to a likely operational error in collecting women's parity data in the 2001 census.*

While the data on children surviving are not directly relevant to the measurement and assessment of fertility, tentative analyses of the children surviving data in the 2001 census suggest that too great a proportion of a woman's children ever born are reported to be still living. Applying Occam's razor, the simplest explanation for both this phenomenon, and the apparent declines in the average number of children born is that women in Census 2001 predominantly reported the number of their living children as their total number of births. This error would explain both the low parities at older ages, as well as the higher ratio of children born to children surviving that is apparent in the 2001 census data relative to that collected in 1996.

2.6 Conclusions

This section has discussed the quality of the fertility data in the 2001 census in greater detail than is usually the case with attempts to estimate rates of fertility from these types of data. Many South African researchers do not even consider the question of data quality, and even among those that do, some confine their consideration solely to whether or not plausible rates can be derived from the data (e.g. Udjo 2003a). It is logically absurd to assert that the production of 'reasonable' rates implies that the data underlying the derivation of those rates are themselves reasonable.

In any event, this section has provided ample evidence that no 'direct' method of estimating fertility levels should be applied to these data given the extent of the errors, inconsistencies and biases inherent in the data. (We do examine the direct estimates in section 6 and show some clear inconsistencies between the directly and indirectly observed estimates). The material presented here should be regarded as a cautionary tale warning against blind application of techniques relying on the 'reasonableness' of one's answers, without first interrogating the quality of the data. The next section derives estimates of fertility from the 2001 census, taking into account the data errors and problems identified here.

3 Derivation of fertility estimates from the 2001 census

As set out in the previous section, the data on fertility collected in the 2001 census offer several challenges to the demographer in the determination of reliable, plausible and useful estimates of fertility.

We first derive estimates of the age distribution of fertility by population group. Section 3.2 then adjusts these corrected age distributions using a variant of the Brass P/F technique to derive estimates of the level of fertility as indicated by the 2001 census data. The final three sections further refine the estimates and present fertility rates by population group and for the country as a whole.

3.1 The shape of fertility distributions in Census 2001

The first step in deriving fertility estimates from the 2001 census data is to define two subgroups of women. The first is that defined to have 'good' fertility data, characterised by none of the responses to the following questions being subject to imputation or hotdecking:

Total children ever born
Total children surviving
Last child born's date of birth (day, month and year not corrected)
Last child born's vital status and sex

All other women were placed in the second group. This marker was applied to all women who were aged 12 or older, but less than 50 on 10 October 2001. As indicated in Table 2.1, just under half of all 'eligible' women had complete and acceptable answers to all these questions.

With the exception of the impossible dates of birth (that had been accepted without correction by the editing procedures), a woman's exact age on the census date was determined from the difference between the census date and her date of her birth, and then grouped into quinquennial age bands. The distribution of the numbers of women in this fashion by age group and population group, and that of women from the original data files based on recoded population group and age were investigated and shown to not differ in any material fashion from each other. However, one significant difference was observed: mother's age at birth of her last child was calculated using the full date of birth of both mother and child. By contrast, the editing rules applied by Statistics South Africa were less precise, relying only on age of mother and age of last born child to determine age of mother at birth. The edit rules used note this error and recommend that "for Census 2006, a more accurate calculation can be made by comparing full dates of birth instead of just ages" (Statistics South Africa 2003:303), ignoring the fact that it is possible to do this using the 2001 data.

It is to be expected that younger women would answer the questions less well than older women, since the forces that give rise to the need for an el-Badry correction would again be in play here. Similarly, women with a great many children are less likely to be included in the first group, since there is a greater likelihood of an error (in recall, or enumeration) being made where more children are involved. The former effect hugely outweighs the latter. The reported average parities of women in the first group (as defined above) were noticeably higher than for women in the second group.

3.1.1 Estimated number of births in the year before the census

The estimated number of births to women in the year before the census, classified by age of mother at date of child's birth, and grouped by the reference group are shown in Table 3.1 and Table 3.2.

Table 3.1 Estimated numbers of births in the twelve months before Census 2001 where neither imputation nor corrections were applied (Reference group 1), by population group and age group of mother at birth

	African	*Coloured*	*Indian*	*White*	*Total*
12-14	2,017	186	7	21	2,232
15-19	75,226	8,092	634	1,270	85,221
20-24	121,103	13,610	2,539	5,457	142,709
25-29	117,426	14,794	3,819	12,066	148,103
30-34	86,052	11,076	2,237	9,808	109,172
35-39	52,904	5,542	836	3,271	62,553
40-44	18,810	1,420	157	646	21,033
45-49	4,051	232	30	123	4,436
Total	**477,588**	**54,951**	**10,258**	**32,661**	**575,458**

Table 3.2 Estimated numbers of births in the twelve months before Census 2001 where either imputation or corrections were applied (Reference group 0), by population group and age group of mother at birth

	African	*Coloured*	*Indian*	*White*	*Total*
12-14	8,129	1,110	72	313	9,624
15-19	106,131	12,920	835	2,610	122,496
20-24	120,914	11,866	1,926	4,677	139,383
25-29	93,060	9,102	2,231	7,299	111,693
30-34	63,064	6,137	1,248	5,354	75,803
35-39	40,796	3,330	447	1,978	46,551
40-44	17,691	1,206	142	526	19,566
45-49	6,032	401	49	173	6,655
Total	**455,818**	**46,071**	**6,951**	**22,930**	**531,771**

While the total reported numbers of births (approximately 1.1 million per annum in 2001) is close to that which would be expected, the number of births based on reliably reported data runs at about half the estimated total.

This is not dissimilar to the situation that pertained in the 1996 census data where only approximately one half of births occurring in the 12 months before the census were reported with any degree of reliability (Moultrie and Timæus 2002, 2003).

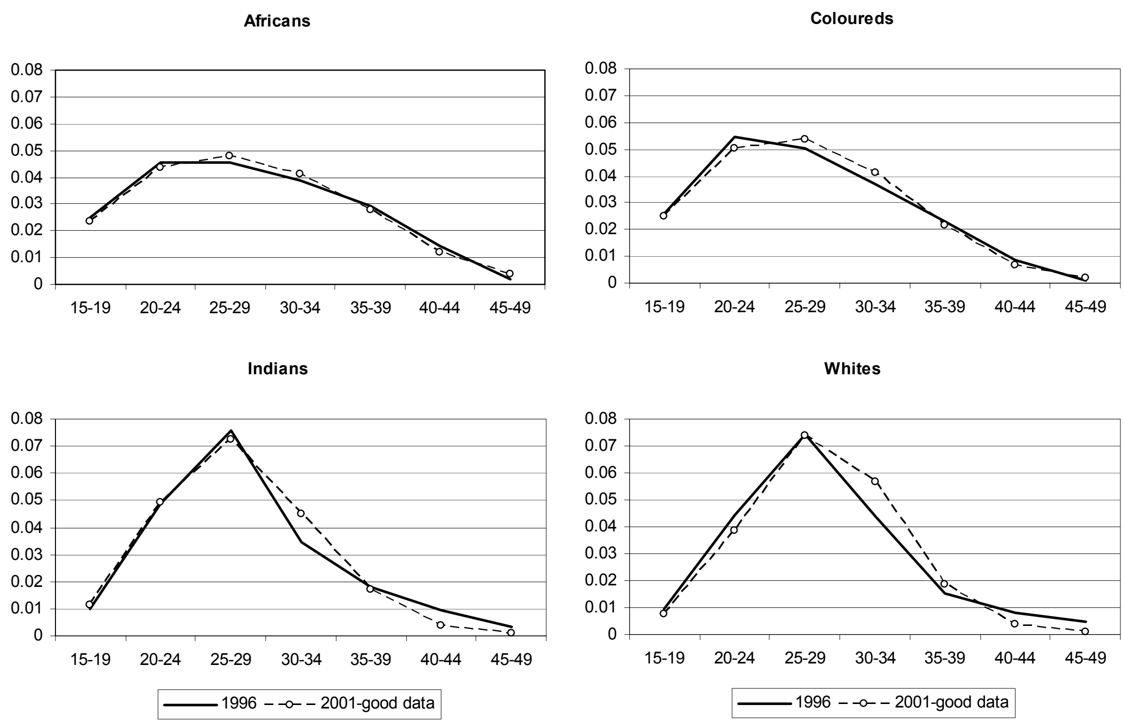
Due to the extent of imputation and hotdecking of the data, and our lack of confidence in the reliability of the methods adopted, our best estimate of the shape of the fertility distribution is to derive age-specific fertility rates by dividing the number of uncorrected births in the 12 months before the census by the (total) number of women enumerated. In effect, this makes the assumption that the data pertaining, and edits rules applied, to adults are reasonable, while those relating to children are not. Further, since dates of birth of women are known (or assumed) to be correct[5] and the exact date of recent births is known (since only the 'reliable' data including non-imputed birth dates for last born children are used), it is preferable to work with the known (to the nearest day) age of mother at the birth of her last child (if born in the twelve months before the census), and the age of women six months before the census who have not. Thus, the unadjusted estimates of fertility relate to the period centred on average six months before the census[6].

First estimates, which will be used to estimate both the shape and the level of fertility by population group, are shown in Table 3.3.

Table 3.3 Initial estimates of fertility from the 2001 census data, by age and population group

Age group	*African*	*Coloured*	*Indian*	*White*	*Total*
12-14	0.001	0.001	0.000	0.000	0.001
15-19	0.037	0.039	0.012	0.007	0.034
20-24	0.068	0.078	0.051	0.038	0.066
25-29	0.075	0.084	0.075	0.073	0.076
30-34	0.065	0.065	0.047	0.056	0.063
35-39	0.044	0.034	0.018	0.019	0.039
40-44	0.019	0.010	0.004	0.004	0.016
45-49	0.006	0.003	0.001	0.001	0.005
Total	**1.57**	**1.56**	**1.03**	**0.99**	**1.50**

Figure 3.1 Standardised fertility distributions, 1996 and 2001 censuses by population group



As would be expected, these estimates of the level of fertility are in and of themselves implausibly low and inconsistent between population groups. However, there are immediate and

[5] Women with illogical birth dates (e.g. 31 September) have been dropped from the analysis.

[6] This is not an unreasonable assumption. Investigation of the (unimputed and unedited) birthdate data used to derive the births in the year before the census shows that these births occurred, on average, on 14 April 2001, only four days off being centred exactly half a year before the census date.

substantive grounds for asserting that the remnant data used in the derivation of these estimates nevertheless have the correct shape: the fertility distributions by population group from the 1996 census are remarkably consistent with those estimated from the 1998 DHS (Moultrie and Timæus 2003). While not constant, the age distribution of fertility usually does not change rapidly over time. As can be seen from Figure 3.1 the distribution of fertility indicated by the data chosen for use from the 2001 census is essentially similar to that from 1996, particularly for the largest of the population groups.

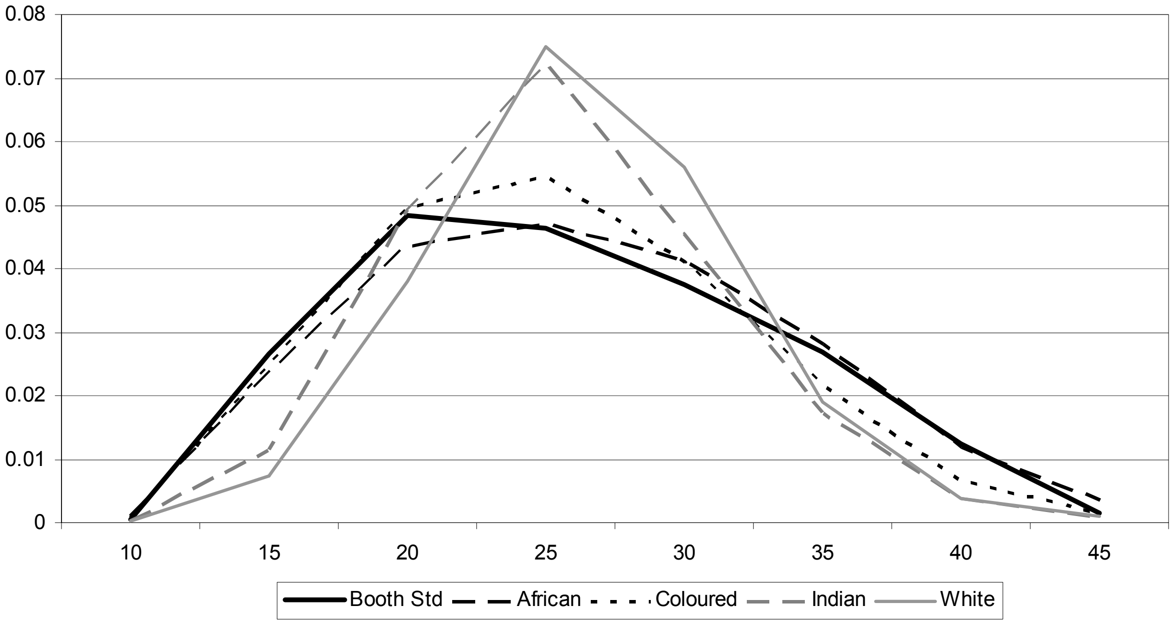
The relative consistency of the fertility schedules at this stage of evaluation is an important feature of the approach adopted.

3.1.2 Relational Gompertz models

The Relational Gompertz model was first set out by Brass in the early 1970s, but the principal development was the extension of the model by Zaba (1981), who used a standard fertility schedule derived by Heather Booth in her PhD thesis, and published subsequently (Booth 1984).

Several demographers working in the region apply the Relational Gompertz model with the Booth standard to all South African population groups. That this is wrong is evident from Figure 3.2.

Figure 3.2 Comparison of fertility schedules from the Booth Standard, and standardised schedules by population group, 2001 census



With the exception of the African population (and to a lesser extent Coloureds), the fertility schedule implicit in the Booth Standard is in no way representative of the fertility of other population groups. Hence, use of this schedule will distort the results of those population groups by constraining the resultant fertility schedules to be more similar to the standard than they in fact are.

Booth herself stresses the inappropriateness of the Relational Gompertz model for use in low fertility populations:

The standard is chosen to be typical of populations with high fertility because the model is intended to be used to detect the kinds of errors that are found in data from such populations" (Booth 1984: 496).

She argues further that

The standard pattern of fertility was developed from a set of high-fertility schedules produced from Coale and Trussell's model … The model is thus designed to fit better in the tails to distributions with a substantial proportion at young or old ages than to distributions with an insignificant proportion in the tails (Booth 1984: 498-9)

While it is common cause that teenage fertility in South Africa is high (and according to the estimates presented here not showing significant signs of decreasing – an issue discussed further in section 7.1), evidence from earlier censuses suggests that old age fertility is not particularly high in any population group. A further reason not to use the Relational Gompertz model in the case of the 2001 data is that the model requires, as an input, parity data by age group. As has already been discussed, the parity data from the 2001 census data are inherently implausible at older ages and separate investigations (requiring estimates of the level and distribution of recent fertility) are required to establish the age (if any) up to which the parity data are credible.

Thus, even under optimal conditions, a Relational Gompertz model would be recommended for use with only the Coloured and African data, and certainly for neither Whites, nor Indians/Asians, nor the country as a whole due to the strong heterogeneity of the fertility distributions for the different population groups[7].

An advantage of the Relational Gompertz model, though, is that in one variant it can correct for errors in age patterns and levels of fertility, or errors in children ever borne data more-or-less independently of each other. However, where both the level of fertility and the average parity data are too low, the Relational Gompertz model is of less use, since it will determine average parities from the level of fertility, or vice versa, and the estimation becomes, as a result, circular.

3.2 Correcting the level of fertility in South Africa: The Brass P/F method, and its variants

It still remains, however, to correct for the level of fertility, even if the shape of the fertility distribution calculated from a subset of the 2001 census data is deemed to be correct. The straight use of a Relational Gompertz model may be of use here, except for the fact that it uses reported parities to determine the level of fertility, and as yet it is uncertain as to which parities are accurate enough to use. An alternative is to use a variant of the Brass P/F Method (United Nations Department of International Economic and Social Affairs 1983) which does not require parties above a certain age. The Brass P/F method only requires parity data up to, at most, the 30-34 age group. Parities after this age are not used in the estimation of fertility rates, while the Relational Gompertz model is affected by the parity data of all age groups included in the model – in most cases extending beyond age 35. Thus, both methods rely to an extent on the accurate reporting of average parities to correct the level of fertility, although the Relational Gompertz

[7] Although, in practice, given the numerical dominance of African and Coloured births, the error is not likely to be material.

model typically uses average parities for age groups beyond that required by the Brass P/F method.

3.2.1 Correction of parity data to estimate fertility levels

In any event, substantial changes must be made to the 'unadjusted' average parity data presented in Table 2.12 prior to our use of these data in a method that seeks to elucidate the 'true' level of fertility from women's reported parity. To this end, it is our contention that the edit rules applied to the census data should not seek to make adjustments to the data where established demographic alternatives exist. We have already presented and discussed the application of an el-Badry correction to flawed parity data. Here, we set out the corrections made – *ab initio* – to the parity data to derive useful estimates of the recent level of fertility in the country. Subsequent sections detail the process whereby 'corrected' estimates of older-age parities were arrived at, and show that these revised estimates are largely consistent with the 1996 census parities on the assumption of constant rates of fertility change (i.e. linear decline) between 1996 and 2001.

The same two reference groups as defined in section 3.1 were used, although the treatment of the raw (i.e. with no edits whatsoever) parity data for each group of women was fundamentally different. The parity data for women whose fertility data required some cleaning or editing were treated as missing for the purposes of estimating average parities, thereby requiring the implicit assumption that the average parities of those with missing data (after application of an el-Badry correction) were consistent with those of women without missing data. The effect of this approach is drastic, as the resultant proportions of estimated 'true parity not stated' women is exceedingly high (running to approximately one third of all women). Despite the severity of the adjustment applied, the average parities resulting from the correction bear a close resemblance to the average parities calculated from the el-Badry adjusted raw parity data presented in Table 2.12, as can be seen in Table 3.4.

Table 3.4 Comparison of estimated average parities after application of an el-Badry correction to both the raw data and data treating all imputed/hotdecked data as missing

	Africans		*Coloureds*		*Indians/Asians*		*Whites*	
	Raw	*Corrected*	*Raw*	*Corrected*	*Raw*	*Corrected*	*Raw*	*Corrected*
15-19	0.16	0.12	0.13	0.11	0.05	0.03	0.04	0.02
20-24	0.67	0.62	0.62	0.57	0.32	0.28	0.24	0.20
25-29	1.38	1.37	1.30	1.26	0.99	0.95	0.79	0.75
30-34	2.23	2.24	1.94	1.92	1.73	1.69	1.42	1.38
35-39	3.01	3.02	2.42	2.40	2.20	2.17	1.84	1.79
40-44	3.51	3.48	2.72	2.69	2.40	2.37	2.03	2.00
45-49	3.81	3.74	2.90	2.85	2.48	2.48	2.10	2.13

However, this is not the final adjustment applied to the parity data before it is deemed useable as an input in a P/F or Relational Gompertz model. Work on earlier census and survey data showed that African and Coloured South African women are consistently likely to include their stillbirths in the enumeration of live births (Moultrie and Timæus 2002). Using data from special probing questions in the 1998 DHS, appropriate correction factors were derived for African and Coloured women and were first applied to the 1996 census data. There is no *prima facie* reason why the 2001 data should not have been subjected to the same errors, especially given the much poorer quality of the data collected in the more recent census. The magnitude of the still births adjustment increases with mother's age, and occasions an approximate 2.5 per cent reduction in the estimated mean parity of older Coloured women, and a reduction of approximately 4 per cent among African women. Our best estimates of average parities, then, from the 2001 census, are shown in Table 3.5. Note, however, that the problem of inconsistency in average parities with the data from the 1996 census has not yet been resolved.

Table 3.5 Comparison of estimated average parities after application of an el-Badry correction to both the raw data and data treating all imputed/hotdecked data as missing, correcting for still births

	Africans		*Coloureds*		*Indians/Asians*		*Whites*	
	Raw	*Corrected*	*Raw*	*Corrected*	*Raw*	*Corrected*	*Raw*	*Corrected*
15-19	0.16	0.12	0.13	0.11	0.05	0.03	0.04	0.02
20-24	0.67	0.61	0.62	0.56	0.32	0.28	0.24	0.20
25-29	1.38	1.34	1.30	1.23	0.99	0.95	0.79	0.75
30-34	2.23	2.18	1.94	1.87	1.73	1.69	1.42	1.38
35-39	3.01	2.94	2.42	2.34	2.20	2.17	1.84	1.79
40-44	3.51	3.37	2.72	2.62	2.40	2.37	2.03	2.00
45-49	3.81	3.59	2.90	2.79	2.48	2.48	2.10	2.13

Note: The 'Raw' data are the same as those reflected in the first column of Table 3.4 and the last column of Table 2.12. The data for Indians/Asians and Whites are the same as in Table 3.4, as no still births adjustment was applied to data for these two groups. They are presented here again for consistency and completeness' sake.

These estimates of average parity by age- and population group are used to estimate the current level of fertility in South Africa at the time of the 2001 census as described in the next section. Final, and corrected, estimates of average parities are derived in section 5.

3.3 Application of the Feeney P/F method

A frequent criticism levelled at the Brass P/F ratio method is that it was devised for use where fertility could be assumed to be constant over time. This is clearly not the case in South Africa. However, Feeney (1998) has shown that by reconceptualising the interpretation of the method but not changing the method in any material way, it is also applicable to situations where fertility is falling. Feeney makes use of an observation by Norman Ryder (1983), that "the mean number of children born to a cohort approximates the period total fertility rate at the time this cohort was at its mean age at childbearing" (Feeney 1998:3). The implication of Ryder's observation is that a period estimate of total fertility can be gained by applying the P/F ratio for the mean age of childbearing, irrespective of whether fertility has been constant or declining, to the entire fertility schedule in order to derive estimates of fertility that preserve the same distribution as the unadjusted estimates, but correct for the level of fertility implied by the parity data of younger women.

The P/F ratios used and mean ages of childbearing in each population group are shown in Table 3.6. With the exception of Indian/Asian women, the means of the fertility schedules have been remarkably stable between the two censuses, while even the magnitude of the adjustments to be applied to the data for each population group in either census are substantively similar.

Table 3.6 Mean ages of childbearing and P/F ratios used to correct the level of fertility by population group, Census 1996 and Census 2001

Year		*African*	*Coloured*	*Indian/Asian*	*White*
1996	Mean age of childbearing (MAC)	28.8	27.6	28.6	28.9
	P/F at MAC	2.05	1.53	2.02	1.65
2001	Mean age of childbearing (MAC)	29.0	27.8	28.1	28.9
	P/F at MAC	1.94	1.54	1.91	1.83

While the mean age of childbearing can vary, possible divergence over the range documented by Udjo (2003b) makes no significant difference to the estimates of fertility derived.

Further, since the mean age of childbearing is sufficiently low, using the Feeney approach to determine the level of fertility avoids the need to use clearly deficient data on parity at older ages. Although in all four population groups, the mean of the fertility schedule lies in the 25-29 age group (centred on 27.5), the use of the parity data for the 30-34 age group (centred on 32.5) is required to estimate the P/F ratio at the mean age of childbearing. However, in all cases the means are much closer to the lower bound used in the interpolation than to the upper, meaning that the results can only be marginally affected by errors in the reported parities at ages 30-34.

Applying the P/F ratios in Table 3.6 to each of the population- and age group specific fertility schedules (shown in Table 3.3) gives our best estimates of fertility that can be derived from the 2001 census data. These estimates, which apply to the year centred at the point six months before the census date (i.e. April 2001), are shown in Table 3.7.

Table 3.7 shows that fertility levels and patterns in South Africa still differ in crucial ways by population group. Fertility levels are highest among Africans, followed by Coloureds, while those of Indians/Asians and Whites are below two children per woman. Importantly too, the age distributions of fertility differ significantly between population groups. Those of Indians/Asians and Whites are strongly concentrated around the mean of the fertility schedule, with low levels of fertility after 35 and before 20. On the other hand, African women's fertility and to a lesser extent that of Coloured women is more evenly spread across the age range from 20 to 40.

Table 3.7 Final estimates of fertility from the 2001 census data, by age and population group

Age group	*African*	*Coloured*	*Indian/Asian*	*White*	*National*
15-19	0.071	0.060	0.022	0.014	0.065
20-24	0.132	0.121	0.097	0.070	0.126
25-29	0.145	0.129	0.144	0.134	0.143
30-34	0.125	0.100	0.089	0.103	0.120
35-39	0.085	0.052	0.034	0.034	0.075
40-44	0.037	0.016	0.007	0.007	0.030
45-49	0.012	0.004	0.002	0.002	0.010
Total	**3.04**	**2.41**	**1.98**	**1.82**	**2.84**

3.4 Estimating the level of fertility nationally

The estimation of national levels of and trends in fertility is not entirely straight-forward. As Moultrie and Timæus (2002, 2003) observed in relation to the 1996 census data, two approaches could be adopted in arriving at an estimated national level of fertility. The first would be to apply the methods and corrections outlined above to the data pertaining to the national population. The second would be to weight the population- and age-specific fertility rates in an appropriate way to arrive at a national estimate. The second approach is preferable, because it accommodates the heterogeneity of the data by population group that the first cannot accommodate directly. With the widely divergent levels and patterns of fertility by population group, the assumptions required in adopting the first approach are – in the view of the authors – unjustifiable.

The resulting estimates of national fertility rates are shown in Table 3.7. The overall level of fertility in the country in 2001 is estimated to be 2.8 children per woman.

3.5 Trends in fertility in South Africa, 1996-2001

The estimates derived here are generally consistent with those produced using earlier census and survey data. In his doctoral thesis, Moultrie (2002) used data from the 1970 and 1996 censuses and the 1998 South Africa Demographic and Health Survey to assess the trends in South African fertility over an extended period of time, and sought to explain the particular pattern of slow and gradual decline in fertility by reference to the social, economic and institutional effects and legacies of apartheid policies. The results from the 1996 census were consistent with that trend, and the trend has continued over the five years between the 1996 and 2001 censuses.

Comparisons of the rates from 2001 with those from 1996 show a rapid decline in fertility in all population groups in the five-year period between the 1996 and 2001 censuses. The level of fertility among Indians is estimated to have fallen by almost 20 per cent in this period. African fertility has fallen by 12.9 per cent, White fertility by 10.3 per cent, and amongst Coloureds by 8.6 per cent. In two of the four population groups, fertility levels are below the 2.1 children per woman benchmarked as 'replacement level fertility', although the true measure of replacement level fertility is affected by the level of childhood mortality. The levels of Coloured fertility continue to fall, and are now only slightly above replacement level, while the fertility among African women is close to three children per woman, a fall of almost half a child per woman in the five years before the 2001 census.

Previous work in estimating fertility in South Africa has relied on other techniques, which have not been used here. The use of reverse-survival models (Bogue 1993) is contra-indicated here due to the wealth of data available from two separate censuses, not to mention the DHS. Apart from this, reverse survival approaches are heavily dependent on estimates of both child and adult mortality in the years immediately preceding the census. These estimates are not yet available. Further, reverse-survival estimates of fertility can be strongly skewed by differential underenumeration of the population of children and infants by single years of age. Earlier research (Dorrington, Budlender and Moultrie 2003) has already identified a significant undercount of children under the age of 5 relative to the estimated overall undercount of 16 per cent. While estimates of the extent of the undercount by single years of age for people older than ten can be derived from a comparison of the two censuses, estimates of the undercount of young children cannot be obtained in the same way, making the estimation of that undercount circularly dependent on the estimate of the level of fertility in the five years preceding the census. A variant of the reverse-survival approach proposed by Cho, Retherford and Cho (1986) that makes use of complex algorithms to link children to their mothers has been applied to the South African population by Sibanda and Zuberi (1999). However, an inherent danger of this approach is that children may inadvertently be linked to their grandmothers or other older women, thereby inflating older-age fertility estimates. Relative to those produced by Moultrie and Timæus (2002, 2003), the estimates produced by Sibanda and Zuberi are orders of magnitude higher at older ages.

Figure 3.3 shows the trend in South African fertility from the middle of the twentieth century. In addition to the two estimates derived from the most recent censuses, the figure also shows the 'official' estimates, given in Mostert *et al* (1998), covering the period up to 1994, as well as the reverse-survival estimates derived from the 1996 census data. As can be seen, the most recent estimates are the plausible continuation of a long-term trend that began in the mid-1960s. The reverse-survival estimates are included as a check on the official data – the variability in these estimates in the 1960s from one year to the next is a direct consequence of differential undercounts of the numbers of children at individual ages: the extent of the undercount of infants (estimated at 26 per cent) and very young children in the 1996 census means that the most recent estimates from the application of a reverse-survival approach are inherently implausible, and are not shown. The comparability of these estimates, together with those produced by Sibanda and Zuberi are discussed at length by Moultrie and Timæus (2002, 2003).

Figure 3.3 Trends in national fertility, 1945-2001

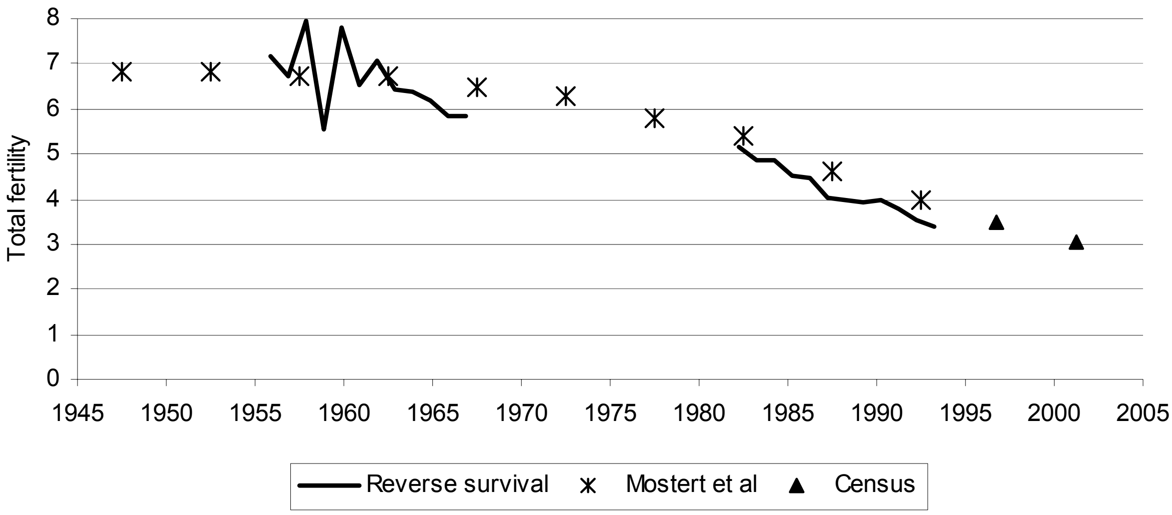
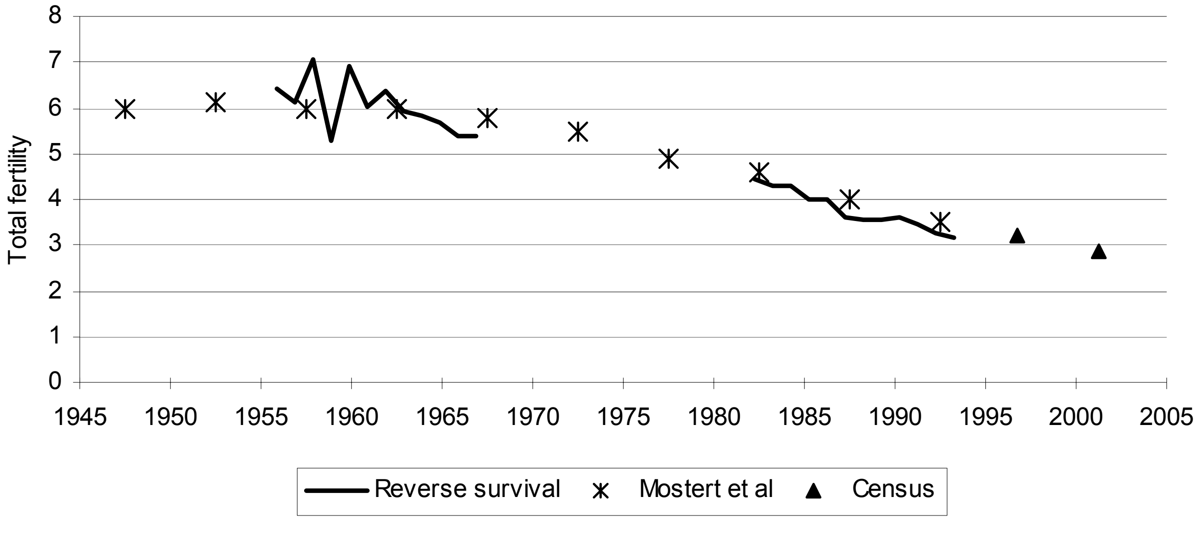


Figure 3.4 shows the same data for African women, and again confirms the ongoing slow decline in African fertility previously identified and discussed.

Figure 3.4 Trends in African fertility, 1945-2001


8
7
6
5
4
3
2
1
0
Total fertility
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
Reverse survival
Mostert et al
Census

4 Provincial fertility in South Africa

Earlier research on fertility in the country has struggled to produce estimates of fertility disaggregated by population group and mothers' province of residence simultaneously because, with the exception of the African population, the breakdown of both births and women by province and population group produces fertility estimates that are too erratic to be of much value. Similar problems pertain to the 2001 census data. For the purposes of this report, we follow the approach adopted by Dorrington, Moultrie and Nannan (2002) by estimating provincial levels of fertility for the African population alone, and using national estimates for Whites, Indians/Asians and Coloureds to derive province-specific fertility rates.

4.1 African fertility by province

The same methods used to estimate African fertility nationally were applied to the data for African women by province. These estimates are shown in Table 4.1. The slightly heterogeneous nature of African fertility by province results in a trivially different estimate of national African fertility when the provincial estimates are aggregated to a national figure (3.08) compared to the 3.04 children per woman cited in the previous section. Which of these figures should be afforded precedence is not easily resolved, as arguments can be constructed as to why either could be the better estimate. However, the small difference suggests that not a lot of effort should be expended on this debate.

Table 4.1 Estimated total fertility for African women by province, 1996 and 2001

	1996	*2001*	*Change (%)*
Western Cape	2.99	2.74	-9
Eastern Cape	4.18	3.46	-17
Northern Cape	3.32	2.61	-21
Free State	3.11	2.60	-16
KwaZulu-Natal	3.73	3.23	-13
North West	3.13	2.85	-9
Gauteng	2.71	2.59	-4
Mpumalanga	3.66	3.23	-12
Limpopo	4.00	3.67	-8
NATIONAL	**3.50**	**3.08**	**-12**

As already indicated, the fertility of African women has fallen by approximately 12 per cent over the five years between 1996 and 2001. This fall is remarkable in several respects. First, the decline in fertility is notable for its spread: in 1996, only two provinces had fertility under three children per woman (the Western Cape and Gauteng), while two provinces (the Eastern Cape and Limpopo) had total fertility levels greater than 4 children per woman. In the intervening five years, no fewer than five provinces show levels of African fertility of less than three children per woman, while three of the remaining four have levels less than 3.5 children per woman. Further, with the exception of Limpopo province, fertility has fallen fastest in those provinces that showed the highest levels of fertility in 1996. The pace of decline has been slowest in Gauteng and the Western Cape. The most plausible explanation for this somewhat surprising finding is that high levels of migration (documented by Dorrington, Budlender and Moultrie (2003)) into these provinces in the intercensal period is responsible for the slow pace of decline. It is commonly accepted (see, for example, National Research Council (2003)) that migrants' fertility tends to trend towards the level of the receiving population. Thus, the fertility of long-term residents of these two provinces may have actually fallen faster than indicated, but the overall pace of decline has been attenuated by the higher fertility of recent in-migrants.

The extremely steep fall in fertility recorded in the Northern Cape is not reliable, since both the African population, and the numbers of births each year are too small to produce robust estimates. This argument cannot be applied to the data for the Free State and the Eastern Cape, and thus the magnitude of these reported declines is surprising and most probably indicative of poor data quality.

The second feature of the data presented in Table 4.1 is the extent of the change in provincial rankings between the two censuses. While Gauteng continues to have the lowest levels of fertility among Africans, among the provinces with highest fertility, fertility in the Eastern Cape (highest in 1996) is now lower than Limpopo (highest in 2001).

A third aspect that deserves comment is the changes observed in the age distribution of fertility in each province between the two censuses. Earlier, it was suggested that one of the reasons for confidence in the results and analysis presented here is that the overall distribution (i.e. after removal of the effect of the decline in fertility) of fertility is fundamentally similar to that identified in the 1996 census (and shown in Figure 3.1).

Fertility is now slightly more concentrated in the 25-34 age group than it was previously, but in essence the age distribution of fertility is similar. This feature of the South African fertility decline is consistent with Caldwell, Orubuloye and Caldwell's hypothesis that in Africa, (unlike Europe) fertility would fall simultaneously in all age groups (Caldwell, Orubuloye and Caldwell 1992).

Of interest, though, is the fact that this similarity is not as uniform when the data are assessed at a provincial level. Figure 4.1 shows the standardised fertility distribution of African fertility by province in both the 1996 and 2001 censuses. With the exception of Gauteng and the Western Cape (and the Northern Cape, for which data are two scant to contradict), there has been little change in the distribution of fertility over the period covered by the two censuses. Fertility in each age group has fallen by roughly the same amount, resulting in lower overall fertility (as indicated in Table 4.1) but similar age distributions of fertility. By contrast, in the Western Cape and Gauteng, there is evidence that the distribution of fertility in 2001 has deviated from that shown by the 1996 data. In each case, the distribution of fertility has become less flat with evidence of a distinct mode in the 25-29 age group. While it is too early to comment on these changes with any degree of certainty, one possible explanation is that in the early and middle stages of African fertility decline, Caldwell, Orubuloye and Caldwell's explanation does indeed hold. However, once fertility has fallen below a particular threshold, the distribution of women's fertility may again change to reflect a more 'typical' low fertility schedule.

4.2 Fertility by province

The fertility data for Coloureds, Whites and Indians/Asians are too sparse to calculate robust and reliable fertility estimates for these populations disaggregated by province.

This is not necessarily problematic, as preliminary investigations of the data indicate that the fertility rates in these groups are fairly consistent within each group across all provinces. As a result, we make the simplifying (and plausible) assumption that fertility patterns and levels in each of those three population groups are constant across all provinces. This is clearly not the case among the African population as was shown above.

Provincial estimates of fertility, across all population groups, were derived in the same way as the national estimate: by calculating the weighted average of age-specific fertility rates in each province, where the weights are the numbers of women in each age and population group. Again, a very small discrepancy in the estimated national rate is apparent.

Table 4.2 Estimated total fertility by province, 1996 and 2001

	1996	*2001*	*Change (%)*
Western Cape	2.60	2.39	-8
Eastern Cape	3.93	3.28	-16
Northern Cape	2.80	2.43	-13
Free State	2.97	2.53	-15
KwaZulu-Natal	3.48	3.04	-13
North West	3.05	2.77	-9
Gauteng	2.55	2.43	-5
Mpumalanga	3.50	3.13	-10
Limpopo	3.94	3.62	-8
NATIONAL	**3.24**	**2.87**	**-11**

The age-specific fertility rates associated with the levels of fertility shown in Table 4.1 and 4.2 are shown in Figure 4.1 and Table 4.3.

Figure 4.1 Standardised fertility distributions for African South Africans by province, 1996 and 2001 South Africa Census

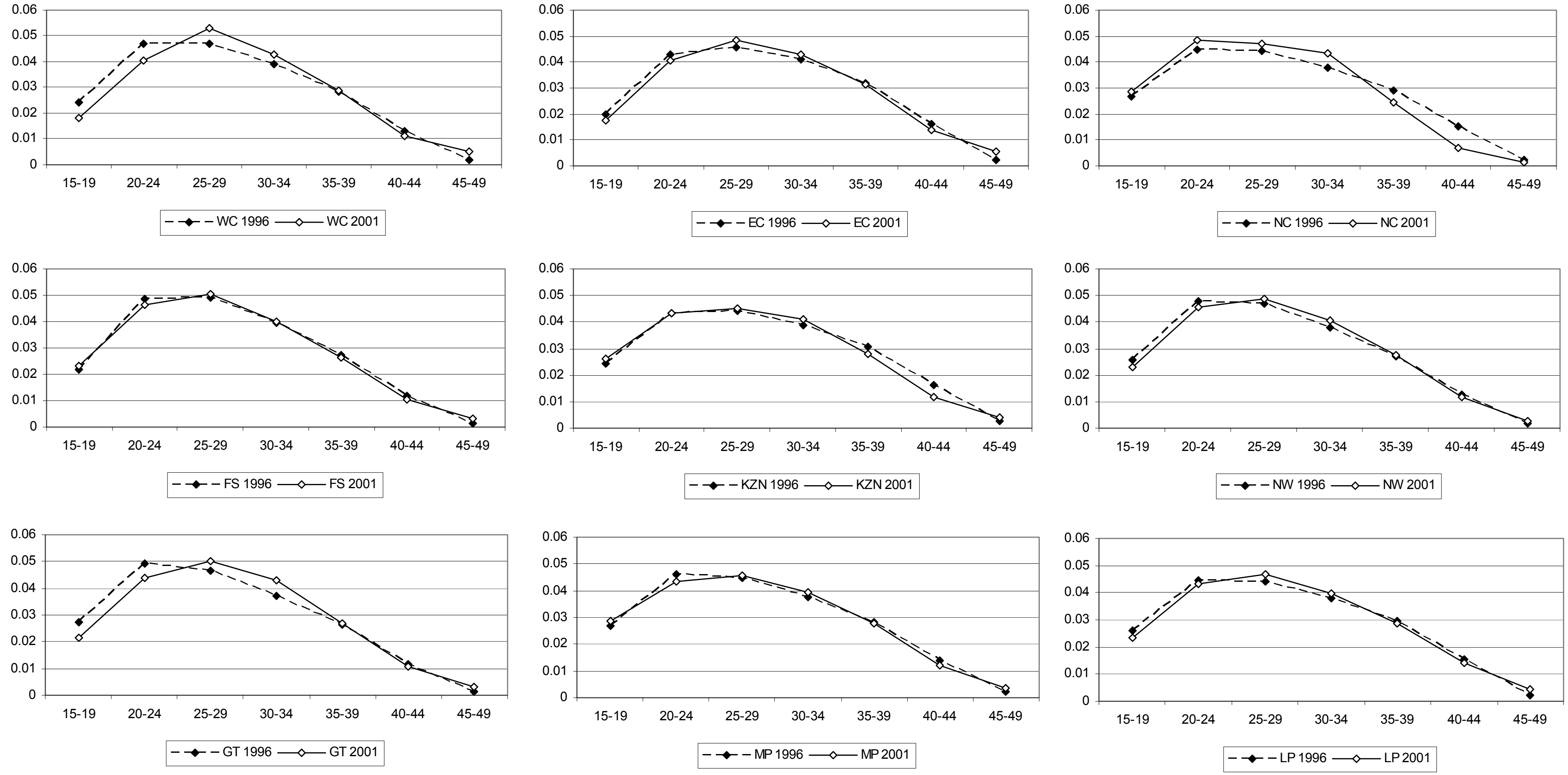


Table 4.3 Estimates of African fertility by province and age group, Census 2001

Age group	*WC*	*EC*	*NC*	*FS*	*KZN*	*NW*	*GT*	*MP*	*LP*	*TOTAL*
15-19	0.050	0.061	0.074	0.060	0.084	0.065	0.056	0.092	0.086	0.073
20-24	0.111	0.140	0.127	0.121	0.140	0.130	0.114	0.140	0.158	0.133
25-29	0.145	0.168	0.123	0.131	0.146	0.139	0.130	0.147	0.171	0.146
30-34	0.118	0.148	0.113	0.104	0.133	0.116	0.112	0.127	0.146	0.127
35-39	0.079	0.108	0.064	0.068	0.090	0.078	0.070	0.089	0.105	0.087
40-44	0.031	0.049	0.018	0.027	0.039	0.034	0.028	0.039	0.051	0.038
45-49	0.014	0.019	0.003	0.009	0.014	0.008	0.009	0.012	0.017	0.013
TFR	**2.74**	**3.46**	**2.61**	**2.60**	**3.23**	**2.85**	**2.59**	**3.23**	**3.67**	**3.08**

Table 4.4 Estimates of fertility by province and age group, Census 2001

Age group	*WC*	*EC*	*NC*	*FS*	*KZN*	*NW*	*GT*	*MP*	*LP*	*TOTAL*
15-19	0.050	0.060	0.061	0.056	0.077	0.062	0.048	0.088	0.085	0.066
20-24	0.110	0.136	0.119	0.118	0.135	0.127	0.108	0.137	0.157	0.127
25-29	0.136	0.162	0.127	0.131	0.145	0.138	0.131	0.146	0.170	0.144
30-34	0.105	0.141	0.105	0.104	0.127	0.114	0.109	0.125	0.144	0.121
35-39	0.055	0.098	0.054	0.064	0.080	0.074	0.061	0.084	0.102	0.076
40-44	0.017	0.043	0.015	0.025	0.032	0.031	0.022	0.035	0.049	0.031
45-49	0.005	0.016	0.003	0.008	0.011	0.007	0.007	0.011	0.016	0.010
TFR	**2.39**	**3.28**	**2.43**	**2.53**	**3.04**	**2.77**	**2.43**	**3.13**	**3.62**	**2.87**

Note: The estimates of White, Coloured and Indian/Asian fertility are assumed to be constant across all provinces, and are assumed to follow the schedules set out in Table 3.7

5 Estimating the true parity of women in South Africa at the time of the 2001 census

This section investigates the parities of women reported in the 2001 census. It has already been established in previous sections that the parity data are seriously flawed and cannot provide a true reflection of women's lifetime fertility. It has been further established that the problems are most pronounced at older ages.

Quite why the data are so poor requires further investigation by Statistics South Africa. We suspect that the answer is multifaceted, and relates to the design of the questionnaire, a lack of time to implement changes identified as necessary in the pilot and possibly inadequate training of enumerators. Sadly, many of these issues were identified in the 1996 census data, and were raised with Statistics South Africa in 2000 in an attempt to avoid a recurrence of the same problems, but with little effect.

However, it is possible to derive estimates of average parities from the 2001 census data using the fertility schedules calculated from the 1996 and 2001 censuses and the average parities of women in the 1996 census.

5.1 Methods and assumptions

Women aged x at the 2001 census date, would have been exactly five years younger (x-5) at the time of the 1996 census. The average parity of women at the time of the 1996 census can be established by single year of age. Likewise schedules of fertility by single year of age can be derived by interpolation from the schedules presented in Table 3.7 and Table 1.1. These single-age schedules can, in turn, be interpolated to provide estimates of fertility in each of the years between 1996 and 2001. Hence, the estimated average parity in October 2001 (0.775 of the way through the year) for a woman aged x last birthday (aged approximately x+½) at the census date, t, is given by

$$\text{CEB}_{x,t} = \text{CEB}_{x-5,t-5} + \sum_{i=1}^{5} \text{ASFR}_{x-i,t-i+0.5}$$

The reason for the half-year lag in the ASFRs is so that the fertility rates used reflect fertility by the end of the census year, and hence are centred on a point half a year before the census date. The derivation of each of the components used is described in detail in the sections that follow.

5.1.1 Average parities in 1996

Data from the 10 per cent public use sample of the 1996 South Africa Census were used to estimate average parities by single-year of age for African women (separately by province, too) and other population groups. Tabulations of the number of children ever born by single year of age of mother (appropriately weighted and restricted for population group and, where required, province) were extracted from these data. Women who were not asked the question on account of their being in an institution were excluded from the tabulation, but missing responses were retained so that an el-Badry correction could be performed on the data. Having adjusted the proportion of women with parity not stated using the el-Badry correction, the revised estimated average parities were further adjusted for the erroneous inclusion of still-births known to have occurred in the 1996 census. This still-birth correction, applied to data for Africans and Coloureds only, was originally calculated on a single-year of age basis, and hence was applied directly to the data.

5.1.2 Single-age fertility rates centred half a year before the census date

The first stage of deriving appropriate fertility rates for the estimation of average parities is to derive single-age fertility rates from both the 1996 and 2001 census data. This was done by straightforward linear interpolation between the fertility rates based on five-year groups of ages calculated and presented previously. The single-year rates for the very oldest and youngest ages were constrained to be zero at 50 and 12, and the sum of the single-year rates (i.e. the total fertility rate) was the same as that derived from the quinquennial schedules.

Due to the different approaches used to derive our best possible estimates of fertility from each census, the resulting single-age fertility estimates derived from the 1996 and 2001 census data apply to October 1996 and April 2001 respectively. Interpolation between these two dates, four and half years apart, is required to estimate single-age-and-year fertility rates centred on April of each year, thereby permitting the assumption that the derived single-age-and-year fertility rates apply to a 12 month period commencing each October.

5.1.3 Average parities in 2001

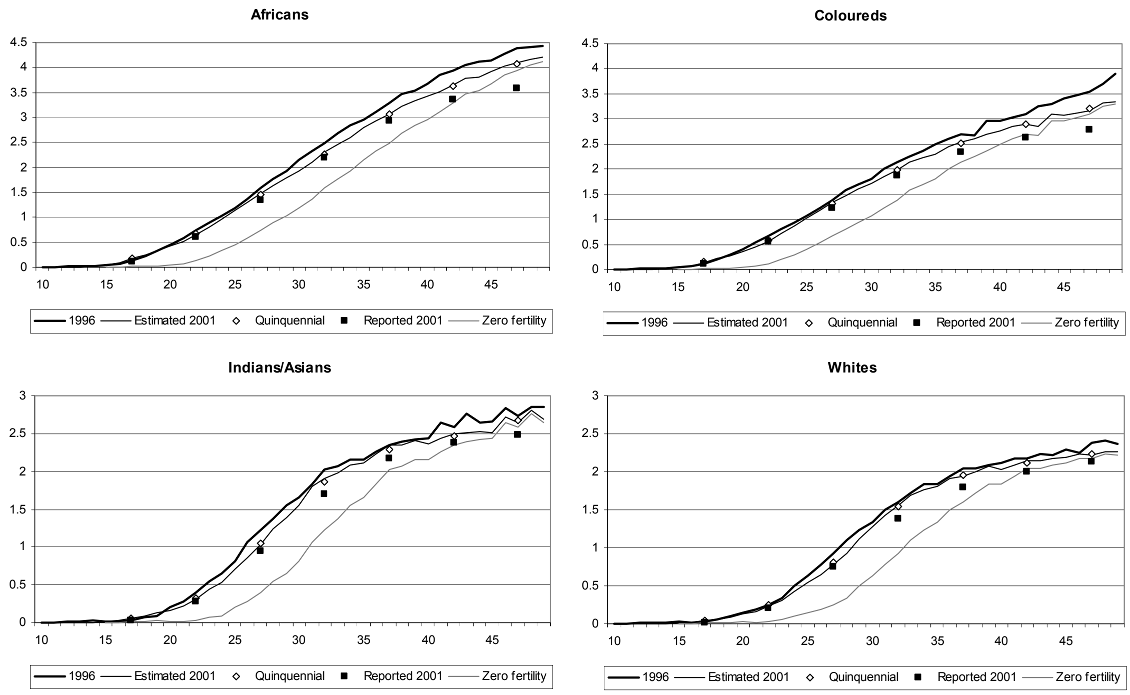
The parity estimates derived in the manner outlined above apply to a single year of age. It is more common, however, to present and interpret data using parities grouped by quinquennial age band. In order to compare the revised indirectly estimated average parities with those estimated directly from the census data (and presented in Table 3.5), an estimate of the grouped average parity was derived using a weighted average, the weights being the number of women of the relevant population group and/or province at each age. Calculating the estimated average parities in 2001 for single years of age using this approach provides an important check on the consistency of the estimated current fertility rates and the reported average parities of women. The results are discussed in the next section.

5.2 Results

The results from the investigation into reporting of average parities are presented in Figure 5.1 and Table 5.1.

Five series of points are presented for each population group. The first, the bold solid line, shows reported parities by age from the 1996 census. The lighter solid line shows the parities that would be expected if both the 1996 parities and the intercensal fertility estimates were correct. The series marked by hollow diamonds shows the same data, but grouped into quinquennial age bands. The fourth series, marked by solid squares shows the best estimate of average parities derived directly from the 2001 census data. The final series, the light grey line is included to show the estimated average parities if we use 1996 parities and assume that there had been no fertility at all in the five years between the censuses. This series acts as a check on the other series, since no estimate for 2001 can fall below it, except as a result of data error.

Figure 5.1 Reported and estimated average parities by population group, 1996 and 2001 census



As discussed earlier, significant data errors occur in the 2001 census data. Reported average parities among women aged 45-49 in all four population groups are lower than is logically possible. However, at younger ages particularly up until age 30, reported average parities and the estimated parities are very close. This correspondence inspires confidence in both the 1996 average parities (which appear, after correction, not to show evidence of the 'overcounting' suggested by Udjo) and hence the current and intercensal estimates of fertility.

Table 5.1 Reported and estimated parities from census data by population group, 1996 and 2001 census

		Africans			*Coloureds*	
Age	*1996*	*2001-reported*	*2001-projected*	*1996*	*2001-reported*	*2001-projected*
15-19	0.164	0.123	0.176	0.136	0.108	0.147
20-24	0.743	0.612	0.670	0.679	0.567	0.573
25-29	1.554	1.340	1.462	1.389	1.233	1.327
30-34	2.496	2.188	2.260	2.106	1.869	1.974
35-39	3.267	2.941	3.069	2.681	2.346	2.509
40-44	3.922	3.372	3.628	3.123	2.627	2.885
45-49	4.333	3.592	4.080	3.597	2.798	3.196

		Indians/Asians			*Whites*	
	1996	*2001-reported*	*2001-projected*	*1996*	*2001-reported*	*2001-projected*
15-19	0.047	0.029	0.064	0.045	0.018	0.041
20-24	0.417	0.282	0.329	0.287	0.200	0.248
25-29	1.198	0.952	1.042	0.918	0.747	0.812
30-34	1.940	1.697	1.858	1.577	1.388	1.540
35-39	2.320	2.171	2.288	1.963	1.796	1.949
40-44	2.609	2.380	2.465	2.154	1.999	2.114
45-49	2.782	2.477	2.673	2.309	2.128	2.233

The increasing deviation of the reported parities (to implausible and illogical levels) also adds to the justification for the method adopted to estimate the overall level of fertility in South Africa from the 2001 census data, since the method relies only on the average parities up to the mean of the fertility schedule. For all four population groups, this implies using the parity data up to the 30-34 age group. However, in each case the mean of the fertility schedule is significantly closer to 27.5 (the mid-point of the 25-29 age group) than it is to 32.5, meaning that less weight is given to the 30-34 parity point than to the (more reliable) point at age 25-29. In any event the data point for 30-34 is reasonable for African and Coloured women, even if less so for Indian and White women.

5.2.1 Estimates of parity by province for African South Africans

A similar exercise to that outlined above was conducted on the data for African South Africans by province. The data are shown in Table 5.2. With the single exception of the Eastern Cape, where the reported parities are a poor fit to those estimated at most ages, the reported average parities in other provinces are consistent up to at least the 30-34 age group. The estimated and reported parities in Gauteng, the North-West and Mpumalanga are consistent up to and including the 35-39 age group. The parity data for KwaZulu-Natal and Limpopo provinces are the best of all the provinces, being inconsistent only in the last age group.

5.3 Conclusions

A major concern at the commencement of this project was that the parity data from the 2001 census were so bad as to be unusable for purposes of analysing trends in and levels of fertility in South Africa.

We have presented here a method of resolving the obvious logical inconsistencies in the parity data from the 2001 census using the fertility and parity data from 1996 together with the estimates of fertility from the 2001 census as presented in the previous section.

The errors documented here have ramifications that extend far beyond the simple calculation of estimates of lifetime fertility. These data are required by almost all indirect techniques used to estimate recent fertility, as well as in the determination of levels and trends in infant and child mortality. Other work for Statistics SA by the same authors (Dorrington, Moultrie and Timæus 2004) presents strong evidence that it will not be possible to derive estimates of child mortality from the 2001 census data.

There can be no doubt that the parity data in the more recent census are severely flawed. While it is impractical to hope that all errors in the parity data collected in a census could be avoided, errors of the magnitude and severity documented here are avoidable. Serious consideration must be given to how to avoid these errors in future South African censuses.

Table 5.2 Average parities for African women by province, 1996 and 2001

	Western Cape			Eastern Cape		
Age	*1996*	*2001-reported*	*2001-projected*	*1996*	*2001-reported*	*2001-projected*
15-19	0.15	0.09	0.16	0.15	0.09	0.15
20-24	0.64	0.49	0.57	0.77	0.52	0.65
25-29	1.36	1.15	1.27	1.72	1.35	1.60
30-34	2.19	1.87	1.99	2.78	2.32	2.58
35-39	2.88	2.53	2.71	3.69	3.18	3.50
40-44	3.58	2.99	3.20	4.34	3.67	4.16
45-49	3.97	3.15	3.71	4.68	3.84	4.54

	Northern Cape			Free State		
	1996	*2001-reported*	*2001-projected*	*1996*	*2001-reported*	*2001-projected*
15-19	0.17	0.13	0.18	0.13	0.11	0.15
20-24	0.73	0.64	0.67	0.64	0.53	0.57
25-29	1.52	1.31	1.39	1.44	1.20	1.32
30-34	2.30	2.02	2.16	2.29	1.96	2.08
35-39	3.06	2.59	2.81	3.01	2.60	2.78
40-44	3.64	2.91	3.38	3.75	3.04	3.31
45-49	4.12	3.21	3.77	4.30	3.34	3.87

	KwaZulu-Natal			North-West		
	1996	*2001-reported*	*2001-projected*	*1996*	*2001-reported*	*2001-projected*
15-19	0.18	0.15	0.20	0.15	0.12	0.16
20-24	0.75	0.69	0.71	0.72	0.60	0.63
25-29	1.58	1.44	1.49	1.49	1.30	1.40
30-34	2.58	2.38	2.32	2.29	2.08	2.15
35-39	3.40	3.21	3.20	3.03	2.76	2.81
40-44	4.11	3.67	3.79	3.75	3.19	3.36
45-49	4.54	3.86	4.30	4.26	3.50	3.88

	Gauteng			Mpumalanga		
	1996	*2001-reported*	*2001-projected*	*1996*	*2001-reported*	*2001-projected*
15-19	0.14	0.10	0.16	0.19	0.17	0.21
20-24	0.66	0.52	0.58	0.81	0.72	0.75
25-29	1.32	1.16	1.25	1.69	1.49	1.56
30-34	2.05	1.85	1.91	2.65	2.38	2.41
35-39	2.60	2.42	2.51	3.48	3.16	3.24
40-44	3.04	2.72	2.88	4.20	3.55	3.85
45-49	3.33	2.87	3.17	4.66	3.76	4.36

	Limpopo		
	1996	*2001-reported*	*2001-projected*
15-19	0.18	0.14	0.20
20-24	0.83	0.74	0.78
25-29	1.72	1.59	1.67
30-34	2.90	2.59	2.54
35-39	3.73	3.54	3.56
40-44	4.43	4.09	4.17
45-49	4.69	4.30	4.61

6 Estimates of fertility calculated directly from the edited data

Other demographers who have had access to the data have also calculated fertility rates from the 2001 census data. However, as argued earlier, simply because it is possible to derive what appear to be plausible estimates from a set of data does not validate the data themselves. Unfortunately, in this regard, the attempts that have been made to date to estimate fertility rates from the 2001 census data have either misapplied the Relational Gompertz model (by assuming that it can be used with all population groups), or used Retherford and Cho's variant of the 'own-child' method (the use of which indicated interpretative shortcomings when applied to the 1996 census data by *inter alia* Sibanda and Zuberi). With the edits applied to the data, it is – at least theoretically – possible to derive estimates of fertility directly from the census data. For reasons set out here, we do not think that much credibility should be attached to the results so obtained, but it is important to contrast the results with those derived in this report

6.1 Direct estimates of fertility

After, but not necessarily as a result of, an earlier intervention by the authors (Dorrington, Budlender and Moultrie 2003) which demonstrated the poor quality of the fertility data in the census and the inadequacy of some of the edits, a team of Statistics South Africa staffers went to the US Census Bureau to work on the edit rules, apparently in an attempt to produce results that would allow users to produce fertility rates directly using the edited data. The results of these edits, as well as the derived variables (for example the number of births in the year preceding the census) necessary to calculate fertility rates have been included in the final release of the census data.

While serious questions have already been raised about the nature of the edits imposed on the data, and their dependence on unreliable variables (such as the MPN) discussed, it is nonetheless important to investigate the fertility rates implied by the edits as they will be calculated by other users of the data, as well as by some demographers, if the fertility data are released as part of the 2001 census data set.

In aggregate, the fertility rates calculated differ only slightly from our estimates. A direct estimate of fertility suggests a total fertility rate of 2.80 children per woman, compared with the 2.84 calculated earlier. Unfortunately, that is where the similarities end.

The results by population group calculated directly from the data, together with the estimates derived earlier, are presented in Table 6.1 and Table 6.2. Other than among African women, the estimated fertility rates differ markedly. The fertility estimates for Coloureds calculated directly from the census data are inherently implausible, placing Coloured fertility levels almost on a par with African fertility. No other research in the country has come close to the suggestion that the two levels are almost equivalent. This is further justification for not relying on the edited fertility data in the census.

Table 6.1 Age-specific and total fertility rates in 2001, best estimate and direct calculations

	National	
Age group	*Moultrie and Dorrington (M&D)*	*Direct*
15-19	0.065	0.074
20-24	0.126	0.128
25-29	0.143	0.132
30-34	0.120	0.110
35-39	0.075	0.072
40-44	0.030	0.033
45-49	0.010	0.012
TFR	**2.84**	**2.80**

Table 6.2 Age-specific and total fertility rates in 2001 by population group, best estimate and direct calculations

	Africans		*Coloureds*		*Indians/Asians*		*Whites*	
Age group	*M&D*	*Direct*	*M&D*	*Direct*	*M&D*	*Direct*	*M&D*	*Direct*
15-19	0.071	0.078	0.060	0.088	0.022	0.022	0.014	0.019
20-24	0.132	0.133	0.121	0.143	0.097	0.084	0.070	0.064
25-29	0.145	0.133	0.129	0.139	0.144	0.120	0.134	0.118
30-34	0.125	0.114	0.100	0.105	0.089	0.078	0.103	0.093
35-39	0.085	0.080	0.052	0.058	0.034	0.031	0.034	0.034
40-44	0.037	0.039	0.016	0.021	0.007	0.008	0.007	0.008
45-49	0.012	0.015	0.004	0.007	0.002	0.002	0.002	0.002
TFR	**3.04**	**2.96**	**2.41**	**2.81**	**1.98**	**1.73**	**1.82**	**1.69**

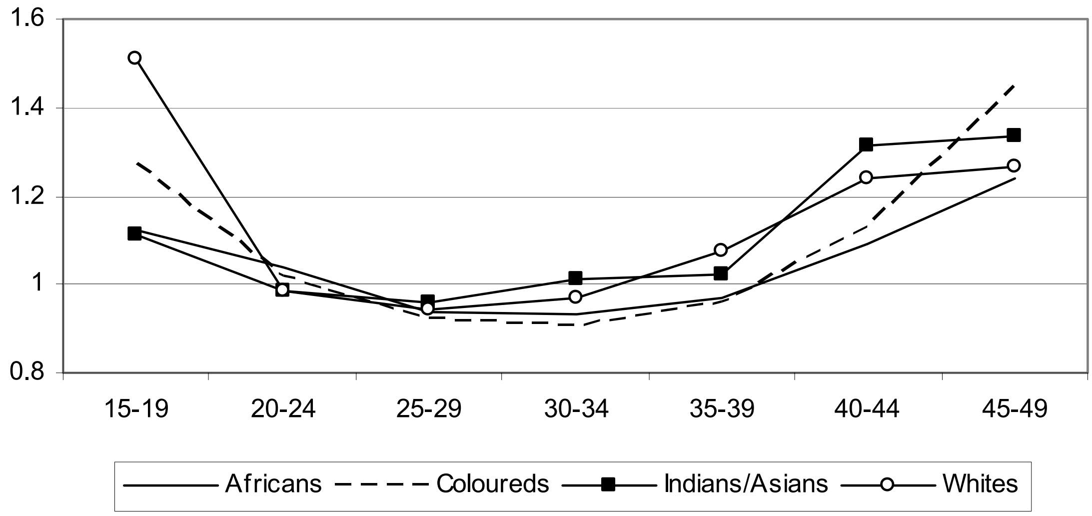
Note: M&D: Moultrie and Dorrington, as presented in previous sections

Likewise, the directly estimated fertility rates suggest an equivalence of Indian/Asian and White fertility levels that is unlikely given the past dynamics of the South African fertility decline (although, admittedly, the margin for error in the derivation of Indian/Asian fertility rates is high due to the small sample size). The age distributions of fertility implied by the directly estimated rates are essentially similar to those estimated using more advanced techniques, except at the oldest (in the case of Africans and Coloureds) and youngest ages, where the directly estimated rates are higher than our best estimates.

Figure 6.1 shows the ratio of the direct to our best estimates of fertility from the 2001 census data by population group. The estimates of fertility have been standardised to remove the effect of the different levels of fertility estimated. If the ratios were all one, this would indicate that the age distributions of fertility implied by each set were identical. A ratio above one indicates the directly calculated rates are higher and vice versa. As can be seen from the figure, the directly calculated rates suggest much higher fertility rates at the youngest and oldest ages.

When looking at the estimated fertility rates by province (Table 6.3), again, a disturbing picture emerges even before consideration of the age distribution of fertility. Among African women, the levels of fertility arrived at by direct estimation and compared with those derived through careful interrogation of the census data agree strongly in seven of the nine South African provinces. However, fundamentally different (and patently implausible) levels of African fertility are suggested for the Eastern Cape and Limpopo provinces. In the case of the former, direct estimation of fertility suggests a level of 2.99 children per woman, compared to the 3.46 estimated indirectly, and the 4.18 estimated in 1996. The directly estimated level of fertility in 2001 is almost 30 per cent lower than that estimated in 1996. Further indications that the fertility data from the Eastern Cape are of particularly poor quality (and that the resultant direct estimates are utterly unreliable) can be found in the implausible parity data collected in the province (see Table 5.2).

Figure 6.1 Ratio of standardised directly calculated and best estimate age-specific fertility rates, by population group, Census 2001



Given the nature of the South African fertility decline, it is unlikely that fertility in one of the highest fertility provinces would fall by nearly 1.2 children per woman when that estimated directly in all other provinces has declined by less than 0.8 of a child.

Table 6.3 Fertility levels from 1996 and 2001 censuses (best and direct estimates), African women

	1996	*Direct 2001*	*M&D 2001*	*Ratio Direct to Best (per cent)*
Western Cape	2.99	2.69	2.74	98
Eastern Cape	4.18	2.99	3.46	86
Northern Cape	3.32	2.66	2.61	102
Free State	3.11	2.68	2.60	103
KwaZulu-Natal	3.73	3.18	3.23	98
North-West	3.13	2.98	2.85	105
Gauteng	2.71	2.66	2.59	103
Mpumalanga	3.66	3.14	3.23	97
Limpopo	4.00	3.19	3.67	87

Source: 1996, Moultrie and Timæus (2002, 2003); M&D 2001, Table 4.1

6.2 Implications

In a perfect world, it would be desirable that any user of the census data with a rudimentary knowledge of the calculation and interpretation of demographic rates should be able to calculate fertility rates directly from the data. In this respect, the efforts made by Statistics South Africa, together with their partners at the US Census Bureau, are commendable. With the data at hand, however, several factors mitigate against this course of action. First, we have expressed concerns about the nature and the quality of the edits applied to the data. Earlier sections of this report have called into question not only the quality of the data themselves, but also errors in the specification of logical edits as well as hotdecks that may result in biases in the data. A flow chart indicating the relative significance of each and every edit applied to the data would go some way to allaying these fears, but the fact remains that having identified one obvious and avoidable error in the specification of the editing rules (for example, the exclusion of higher parity never-married women from the hotdecks), one cannot be certain that other errors have not been made too.

Second, while it would appear that the edits (and recall the odd distribution of births in 2000 and 2001 subjected to editing and hotdecking) result in similar aggregate levels of fertility to those estimates arising from a more robust and theoretically sound approach to the data, this similarity is not sufficient to promote their acceptance. When disaggregated by population group or province (for African women), inexplicable inconsistencies arise. We can be almost certain that Coloured fertility is not almost on a par with African fertility. Likewise, we are sceptical of the suggested declines in African fertility in the Eastern Cape, a decline at variance with all other census and survey data.

Third, while we concur fully with the view of Statistics South Africa that the census can be used for nation-building and should be accessible to all citizens, we question the value of making flawed data available to the population from which will be derived flawed estimates. Furthermore, the level of uncertainty increases with greater degrees of disaggregation and to provide the user with output that suggests that fertility rates can be calculated to any greater degree of disaggregation than by province and population group is to mislead rather than inform.

For these reasons (dissatisfaction with the editing rules and procedures applied to the fertility data; the dangers associated with disaggregating fertility data to spatial levels more refined than province; and the implausibility of some of the results produced by direct estimation) it is our considered opinion and recommendation that the fertility data collected in Census 2001 be released subject to the following:

1. The derived variables that allow direct calculation of fertility rates (e.g. births in the last year) not be released;
2. No edits based on the use of hotdecks as applied to the fertility questions (P20) be admitted;
3. No logical imputation edits which rely directly (or indirectly) on Mother or Father Person numbers be admitted.

7 Conclusions and discussion

This section draws the threads from the preceding sections together, and synthesises the results presented. First, we reflect on the ongoing decline in South African fertility as indicated by the 2001 census data. We then examine one particular aspect of this decline in greater detail, the extent of teenage childbearing, as this has important public policy consequences as well as bears directly on young women's sexual and reproductive health in a time of HIV/AIDS. Finally, we discuss the usability of the fertility data collected in the 2001 census.

7.1 Declining fertility in South Africa

This report has delineated the patterns and recent trends of South African fertility between 1996 and 2001. We have documented in exhaustive detail the quality of the data, and have expressed strong reservations about the processes used to clean and edit the fertility data. Using careful analytical techniques, and paying particular attention to matters of data quality, we have succeeded in deriving what we feel are the most robust estimates of fertility (by province and population group) that can be derived from the 2001 census data. In the process of doing so, we have additionally reconstructed the estimates of women's parity.

Fertility rates in South Africa continue to fall. By any measure, fertility among Whites and Indians/Asians is now well below replacement level. Fertility among Coloured women continues to fall, but more slowly than that of other population groups, while that among African women has declined by almost half a child per woman in the five years between 1996 and 2001. With an estimated fertility rate of marginally over three children per woman among Africans and approximately 2.8 children per woman nationally, fertility levels in South Africa are unquestionably the lowest in sub-Saharan Africa.

The recent decline in South African fertility is a continuation of a process of gradual decline set in motion in the mid-1960s. Looking to the future, the experience of (developed) countries already showing low levels of fertility suggest that further declines in the level of fertility are likely among all population groups. A further factor should be mentioned here, namely the effects of HIV/AIDS. High levels of HIV prevalence are causally associated with decreases in fertility. The vectors through which this association operates are those associated with lower fecundability as a consequence of co-infection with other sexually transmitted diseases and maternal mortality and morbidity. Analysing and understanding the extent to which the ongoing decline in South African fertility is being driven by the spread of HIV is one of the key demographic research areas of the next few years.

A further implication of the spread of the epidemic is that replacement fertility rates will rise to reflect the increase in AIDS-related peri-natal, infant and child mortality.

Finally, we should note here that an important milestone in South Africa's demographic evolution has been passed. According to our best estimates, the number of births per year peaked several years ago, before the 2001 census, and is now declining. This shift, from the number of births increasing at a decreasing rate, to decreasing at an increasing rate is of particular importance for the planning and implementation of developmental initiatives.

7.2 Young adult and teenage pregnancy

A key marker of the success of HIV/AIDS intervention strategies in the country lies in the avoidance of teenage pregnancy. By definition, pregnancy is a possible outcome from unprotected sexual intercourse, which itself places the woman (or her partner(s)) at risk of infection if either she, or her partner(s), are infected with the virus. Given that avoidance of pregnancy can be achieved without attenuating risk of exposure to HIV (through the use of hormonal contraceptives), high levels of teenage pregnancy are likely to be consistent with high levels of exposure to the virus.

A careful and deeper analysis of the fertility data presented earlier paints a gloomy picture of teenage pregnancy in South Africa. On a superficial level, it would appear that the problem of teenage pregnancy is declining in South Africa: fertility rates among women aged 15-19 have fallen between 1996 and 2001, as shown in Table 7.1.

Table 7.1 Fertility rates (births per 1 000 women) among women aged 15-19 by population group, 1996 and 2001 censuses

	African	*Coloured*	*Indian/Asian*	*White*
1996	87.3	68.3	24.1	19.2
2001	72.6	59.7	22.2	13.5
Decline (per cent)	-16.8	-12.7	-7.8	-29.8

In each population group, the absolute level of teenage fertility has fallen, although the rate in respect of African and Coloured women is still high: almost one in 14 African girls aged 15-19 are expected to give birth each year. In addition, since these rates only record live births, the pregnancy rate is undoubtedly considerably higher. Some indication of this can be found in the recent survey of sexual behaviour among young South African women (Pettifor, Rees, Steffenson *et al.* 2004) that showed that 15 per cent of all 15-19 year olds, and no less than 33 per cent of sexually active 15-19 year olds had been pregnant.

However, the data in Table 7.1 does not give the full picture. Examination of the proportion of births born to teenage mothers (Table 7.2) suggests that fertility is more concentrated among Coloured and Indian/Asian teenagers in 2001 than it was in 1996. The proportion of births born to teenage African mothers has remained unchanged, while it has declined among Whites.

Table 7.2 Proportion of births (per cent) born to teenage mothers by population group, Census 1996 and Census 2001

	African	*Coloured*	*Indian/Asian*	*White*
1996	15.9	14.0	5.2	4.5
2001	15.9	14.8	6.2	3.9

Interestingly, when we examine the data for Africans by province (Table 7.3), strong differences are apparent. Among African women, a shocking one in every five births in Mpumalanga, Limpopo and KwaZulu-Natal are to teenage mothers. In four out of nine provinces, the proportion of births to teenage mothers has increased since 1996. Only in the Western Cape and Gauteng has the proportion remained 'low' – at one in ten births. One possible explanation for the lower rates in these two provinces must be that contraception and termination services are more readily available, accessible and subject to less social stigma than in other provinces. This is essentially the same conclusion as that reached by Dickson, Jewkes, Brown *et al* (2003), although they also document women's reliance on private-sector termination facilities, which also are concentrated in these two provinces. These data point to a public health and social crisis. Many teenage mothers have their schooling interrupted, or fail to complete school thereby restricting the potential for their participation in the labour force. Maternal mortality is higher among younger mothers. Also, as suggested above, by definition maternity requires unprotected intercourse, thereby exposing the young mother to infection with HIV.

Table 7.3 Proportion of births (per cent) born to African teenage mothers by province, Census 1996 and Census 2001

Province	*African 1996*	*African 2001*
Western Cape	10.6	10.0
Eastern Cape	16.1	15.3
Northern Cape	16.4	17.3
Free State	13.3	14.9
KwaZulu-Natal	16.3	18.9
North-West	15.9	14.8
Gauteng	11.4	10.1
Mpumalanga	17.4	19.7
Limpopo	20.4	19.0

7.3 Conclusions: Fertility data in the 2001 census

The ability to produce 'plausible' estimates of fertility from census data does not carry with it the implication that the data themselves are either plausible or reasonable. To argue so is meretricious and logically flawed.

The estimates derived here carry other implications with them that have not been discussed in the present report. The most important of these is that the estimated levels of fertility (and the trend in fertility over the five years between the two most recent censuses) are inconsistent with the enumerated population of children under the age of five allowing for our current best estimates of child and infant mortality. This inconsistency is particularly noticeable among Whites and Indian/Asians (where the enumerated population is only seventy percent of that implied by these fertility rates), but is common to all population groups. Although undoubtedly most of this difference is due to errors in the enumeration of young children in South Africa in the 2001 census, accepting the estimates of fertility for Whites and Indians/Asians implies a greater underenumeration in these two population groups. To reject this conclusion, however, would be to suggest that there has been a dramatic (and globally unprecedented) collapse in White and Indian/Asian fertility over the five years prior to the 2001 census.

The nature of the errors in the 2001 fertility data are such that, at best, the data should be used with extreme caution and then only by trained demographers, who are able to spell out the implications of their choices, decisions and methods on any resulting estimates of fertility. In short, the conclusion we come to is that the fertility data collected in the 2001 census are useable only after the utmost care has been taken in the analysis and determination of the data to be used in the estimation of fertility rates. *We would strongly motivate that the derived variables from which estimates of fertility can be calculated directly not be released to the general public, as doing so will lead to the production of divergent, inconsistent and internally flawed estimates of fertility.*

References

Bogue, Donald J. 1993. "Fertility measurement by reverse survival of age-sex distributions," in Bogue, Donald, J., Eduardo E. Arriaga and Douglas Anderton, L. (eds). *Readings in Population Research Methodology*. Vol. 3. New York: United Nations Fund for Population Activities, pp. 11/39 - 11/42.

Booth, Heather. 1984. "Transforming Gompertz' function for fertility analysis: The development of a standard for the relational Gompertz function", *Population Studies* **38**(3):495-506.

Brass, William. 1996. "Demographic data analysis in less developed countries: 1946-1996", *Population Studies* **50**(3):451-467.

Caldwell, John C., Israel O. Orubuloye and Pat Caldwell. 1992. "Fertility decline in Africa: A new type of transition?" *Population and Development Review* **18**(2):211-242.

Cho, L.J., Robert D. Retherford and M.K. Cho. 1986. *The Own-Child Method of Fertility Estimation.* Honolulu: University of Hawaii Press.

Cleland, John. 1996. "Demographic data collection in less developed countries", *Population Studies* **50**(3):433-450.

Dickson, Kim Eva, Rachel K. Jewkes, Heather Brown *et al.* 2003. "Abortion service provision in South Africa: Three years after liberalisation of the law", *Studies in Family Planning* **34**(4):277-284.

Dorrington, Rob, Debbie Budlender and Tom A. Moultrie. 2003. *Census 2001 Group 3 Report: Are the numbers what we expect?* Pretoria: Statistics Council.

Dorrington, Rob, Tom A. Moultrie and Nadine Nannan. 2002. "Estimates of provincial fertility and mortality 1985-1996.," Paper presented at Annual Conference of the Demographic Association of Southern Africa. Cape Town, 26-27 September 2002.

Dorrington, Rob, Tom A. Moultrie and Ian M. Timæus. 2004. *Estimation of mortality using the South African Census 2001 data.* Cape Town: Centre for Actuarial Research for Statistics South Africa.

el-Badry, M.A. 1961. "Failure of enumerators to make entries of zero: errors in recoding childless cases in population censuses", *Journal of the American Statistical Association* **56**(296):909-924.

Feeney, Griffith. 1998. *A New Interpretation of Brass' P/F Ratio Method Applicable when Fertility is Declining.* http://www.gfeeney.com/notes/pfnote/pfnote.htm. Accessed: 11 January 2000.

Mostert, W.P., B.E. Hofmeyr, J.S. Oosthuizen *et al.* 1998. *Demography: Textbook for the South African Student.* Pretoria: Human Sciences Research Council.

Moultrie, Tom A. 2002. "Apartheid's Children: Social Institutions and Birth Intervals During the South African Fertility Decline, 1960-1998." Unpublished PhD thesis, London: University of London.

Moultrie, Tom A. and Ian M. Timæus. 2002. *Trends in South African fertility between 1970 and 1998: An analysis of the 1996 Census and the 1998 Demographic and Health Survey.* Cape Town: Medical Research Council. http://www.mrc.ac.za/bod/trends.pdf.

Moultrie, Tom A. and Ian M. Timæus. 2003. "The South African fertility decline: Evidence from two censuses and a Demographic and Health Survey", *Population Studies* **57**(3):265-283.

National Research Council. 2003. *Cities Transformed: Demographic Change and Its Implications for the Developing World.* Panel on Urban Population Dynamics, M.R. Montgomery, R. Stern, B. Cohen and H.E. Reed (eds), Committee on Population, Division of Behavioural and Social Sciences and Education. Washington DC: National Academies Press.

Pettifor, Audrey E., Helen V. Rees, A. Steffenson *et al.* 2004. *HIV and Sexual Behaviour Among Young South Africans: A National Survey of 15-24 Year Olds.* Johannesburg: Reproductive Healtth Research Unit, University of the Witwatersrand.

Phillips, Heston, Marius Cronje and Eileen Phoshoko. 2003. "Fertility levels and trends in South Africa: Evidence from the 2001 Census of Population," Paper presented at 2003 Annual Conference of the Demographic Association of Southern Africa. Potchefstroom, 14-16 October 2003.

Ryder, Norman B. 1983. "Cohort and period measures of changing fertility," in Bulatao, Rodolfo A. and Ronald D. Lee (eds). *Determinants of Fertility in Developing Countries.* Vol. 2. New York: Academic Press, pp. 737-756.

Sibanda, Amson and Tukufu Zuberi. 1999. "Contemporary fertility levels and trends in South Africa: Evidence from reconstructed census birth histories," Paper presented at Third African Population Conference. Durban, South Africa, 6 - 10 December 1999. Union for African Population Studies. Vol. 1:79-108.

Statistics South Africa. 2003. *Census 2001: Computer Editing Specifications (30 September 2003).* Pretoria: Statistics South Africa.

Statistics South Africa. 2003. *Computer Edit Specifications for Census 2001 Data Processing Project (18 March 2003).* Pretoria: Statistics South Africa.

Udjo, Eric O. 1998. *The People of South Africa - Population Census 1996: Additional Evidence Regarding Fertility and Mortality Trends in South Africa and Implications for Population Projections.* Pretoria: Statistics South Africa, Directorate of Analysis. http://www.statssa.gov.za/Staticpages/PDF_Whitepapers/udjo.pdf Accessed 4 January 2004.

Udjo, Eric O. 2003. "Is the fertility information from the 2001 South Africa population census useable?" Paper presented at Demographic Association of Southern Africa 2003 Conference. Potchefstroom, South Africa, 14-16 October 2003.

Udjo, Eric O. 2003. "A re-examination of levels and differential *(sic)* in fertility in South Africa from recent evidence", *Journal of Biosocial Science* **35**:413-431.

United Nations, Department of International Economic and Social Affairs. 1983. "The el-Badry Correction for Children Ever Born," in *Manual X: Indirect Techniques for Demographic Estimation, Annex II.* New York: United Nations, pp. 230-235.

United Nations Department of International Economic and Social Affairs. 1983. "Estimation of Fertility Based on Information about Children Ever Born," in *Manual X: Indirect Techniques for Demographic Estimation, Annex II.* New York: United Nations, pp. 27-37.

Zaba, Basia. 1981. *Use of the Relational Gompertz Model in Analysing Fertility Data Collected in Retrospective Surveys.* Centre for Population Studies Research Paper 81-2. London: Centre for Population Studies, London School of Hygiene & Tropical Medicine.