University of Cape Town

CENTRE FOR
SOCIAL SCIENCE RESEARCH

THE SENSITIVITY OF ESTIMATES OF POST-APARTHEID CHANGES IN SOUTH AFRICAN POVERTY AND INEQUALITY TO KEY DATA IMPUTATIONS

Cally Ardington, David Lam,
Murray Leibbrandt and
Matthew Welch

CSSR Working Paper No. 106

Published by the Centre for Social Science Research
University of Cape Town
2005

Copies of this publication may be obtained from:

The Administrative Officer
Centre for Social Science Research
University of Cape Town
Private Bag
Rondebosch, 7701
Tel: (021) 650 4656
Fax: (021) 650 4657
Email: kforbes@cssr.uct.ac.za

Price in Southern Africa (incl. VAT and postage): R 15.00

or it can be downloaded from our website
http://www.cssr.uct.ac.za

ISBN 1-77011-036-4

CENTRE FOR SOCIAL SCIENCE RESEARCH

Southern Africa Labour and Development Research Unit

THE SENSITIVITY OF ESTIMATES OF POST-APARTHEID CHANGES IN SOUTH AFRICAN POVERTY AND INEQUALITY TO KEY DATA IMPUTATIONS

Cally Ardington, David Lam,
Murray Leibbrandt and
Matthew Welch

CSSR Working Paper No. 106

February 2005

Cally Ardington is a Lecturer in the Department of Statistical Sciences, a SALDRU Research Associate at the University of Cape Town and a Visiting Scientist, Africa Centre for Health and Population Studies.

David Lam is a Professor at the Population Studies Center, University of Michigan and a SALDRU Research Associate in residence for 2005 at the Centre for Social Science Research (CSSR).

Murray Leibbrandt is a Professor in the School of Economics at the University of Cape Town and the Director of the Southern Africa Labour and Development Research Unit (SALDRU) within the CSSR.

Matthew Welch is the Deputy Director of the Data First Resource Unit within the CSSR.

The Sensitivity of Estimates of Post-Apartheid Changes in South African Poverty and Inequality to key Data Imputations

Abstract

We begin by summarising the literature that has assessed medium-run changes in poverty and inequality in South Africa using census data. According to this literature, over the 1996 to 2001 period both poverty and inequality increased. In this paper we assesses the robustness of these results to the large percentage of individuals and households in both censuses for whom personal income data is missing and to the fact that personal income is collected in income bands rather than as point estimates. First, we use a sequential regression multiple imputation approach to impute missing values for the 2001 census data. Relative to the existing literature, the imputation results lead to estimates of mean income and inequality (as measured by the Gini coefficient) that are higher and estimates of poverty that are lower. This is true even accounting for the wider confidence intervals that arise from the uncertainty that the imputations bring into the estimation process. Next we go on to assess the influence of dubious zero values by setting them to missing and re-doing the multiple imputation process. This increases the uncertainty associated with the imputation process as reflected in wider confidence intervals on all estimates and only the Gini coefficient is significantly different from the first set of estimated parameters. The final imputation exercise assesses the sensitivity of results to the practice of taking personal incomes recorded in bands and attributing band midpoints to them. We impute an alternative set of intra-band point incomes by replicating the intra-band empirical distribution of personal incomes from a national income and expenditure survey undertaken in the year before each census. Using the empirical distributions increases estimated inequality although the differences are relatively small. We finish our empirical work with a discussion of provincial poverty shares as a policy relevant illustration of the importance of dealing with missing values. Overall our results for 1996 and 2001 confirm the major findings from the existing literature while generating more reliable confidence intervals for the key parameter of interest than are available elsewhere.

1. Introduction

Changes in inequality and poverty are key dimensions of the transformation of any society. Given the twentieth century history of South Africa, these two dimensions of economic well-being and, in particular, their changing racial profiles have been of special interest. One of the important empirical traditions in tracking longer-run South African inequality and poverty changes has made use of records of personal income collected in the national censuses of 1970, 1991 and 1996 (McGrath (1983), Whiteford and McGrath (1994), Whiteford and van Seventer (2000)). In the apartheid era, such empirical work was central to highlighting the destructive impact of racially driven policies on South Africa's non-white groups. In the post-apartheid era, these empirical analyses have taken on additional importance. The size and national reach of the 10 per cent micro sample from the 1996 census made it uniquely suited to deriving a set of district and provincial level poverty profiles that could be used to inform provincial and municipal budgetary allocations for various anti-poverty policies (Babita *et al* (2002)).

In 2004, the ten percent micro-sample from the 2001 census was released. This made it possible to use 1996 and 2001 micro-data to track immediate post-1994 progress in undoing the apartheid legacy. Leibbrandt *et al.* (2004) and Simkins (2005) presented initial results on the changes in the levels and composition of income inequality and poverty between 1996 and 2001 using these data. Whiteford and van Seventer (2000) had documented a high but constant national income inequality for the 1991 to 1996 period. Both Simkins (2005) and Leibbrandt *et al* (2004) showed that this inequality remained high and even took a turn for the worse. As regards racial inequality, between 1996 and 2001, inequality within each race group increased. Formal decompositions showed that this within-group contribution to aggregate inequality increased while the between-group component decreased. This represented a continuation of the trend that Whiteford and van Seventer (2000) had noted for the 1991-1996 period and, indeed, from as far back as 1975.

The poverty analysis of Simkins (2005) and Leibbrandt *et al* (2004) revealed that national poverty worsened over the period, particularly for Africans. This suggested a continuation of the longer-run poverty trend revealed by Whiteford and van Seventer (2000). However, for the 1996-2001 period, the extent to which poverty was measured as increasing was very much dependent on the choice of poverty line. At lower poverty lines, the increase in poverty is significantly more muted than at higher poverty lines. [1]

The rationale behind Leibbrandt *et al* (2004) was to produce comparable empirical results to those of analyses of earlier years, such as that of Whiteford and van Seventer (2000). Such comparability demands that detailed attention be given to replicating data assumptions and methods that were used on the pre-1996 data sets. These methods are not necessarily current best practice and a number of improvements can be considered as soon as the focus switches to building up the best possible analyses of the income data for each of 2001 and 1996 without regard to longer-run comparability.[2] It is certainly possible to undertake a thorough set of imputations for missing values on the personal income variable in the census and to ascertain the sensitivity of key income-based measures of well-being to these imputations. This is the broad task of our paper.[3]

In the 1996 and 2001 censuses, data on personal incomes is gathered by means of a question asking each person in the household 'What is the income category that best describes the gross income of (this person) before tax?' (Statistics South Africa, (1996: 6) and Statistics South Africa, (2001a: 3)).[4] While the

[1] The Leibbrandt *et al* (2004) paper goes on to complement the analysis of the income inequality and income poverty changes with an analysis of changes in access to services. This access-based approach focused on type of dwelling, access to water, energy for lighting, energy for cooking, sanitation and refuse removal. These data on access revealed significant improvements in access between 1996 and 2001. The contrast between these findings and the findings on income serve as important reminders that income is only one of many dimensions to well-being and that non-money metric aspects of well-being are important.

[2] Even from the standpoint of consistency in the way that the census data was collected, there is some justification for an exclusive focus on 1996 and 2001:

> 'In the apartheid years, different approaches were used for enumeration in different areas. In particular, some 'black' areas were "enumerated" by means of estimates from aerial photographs as it was considered too dangerous for enumerators to go door-to-door. The 1996 census was the first attempt to standardise methodology for all areas, and this practice was repeated in 2001' (Cronje and Budlender 2004: 68).

[3] Simkins (2005) makes a promising start down this road. For both 1996 and 2001, a set of decision rules is applied to allocate positive incomes to some adults with missing incomes and to adults with zero incomes that are in households with zero income. These decision rules are overt and replicable. However, they are not anchored in the imputation literature and there is no testing for the sensitivity of results to plausible rule changes.

[4] In both years, the respondent was told that the reference period was 1 October of the previous year until 31 September of the census year. They were also told that this income should include all sources of income including housing loan subsidies, bonuses, allowances such as car allowances, investment income as well as any pension or disability grants. In 1996, the first question of a subsequent household module prompted respondents about "additional money that this household generates and that has not been included in the

broad reach of the census data is its strength, this income data is far from ideal. Cronje & Budlender (2004) highlight one particular weakness; namely, that in both 1996 and 2001, the question on personal income requested an appropriate income band for each person rather than an income value. These bands were not a consistent set of real income categories across the two years. This is especially true at the top end. The highest band for personal income in 1996 was R30 000 or more. This is lower than the real income equivalent of the top three bands in 2001. This incompatibility of income bands in real terms needs to be dealt with in order to compare the data across time.[5] There is no particular subtlety to the decisions that analysts make in this regard and the most that can be asked for is that the decisions are spelled out explicitly and that there is some assessment of the sensitivity of any analysis to alternatives.

A more important but largely unexplored consequence of the fact that personal incomes are recorded in bands is the fact that all those using the income variable for poverty and inequality analysis have to translate the bands into point incomes for each person. The general practice in South Africa has been to attribute the band midpoints to all individuals. This is only one of a number of possibilities and one of the tasks of this paper is to assess the importance of different within-band point income allocation rules.

previous section. (For example, the sale of home grown produce of home-brewed beer or cattle or rental of property about remittance income.)" (Statistics South Africa, 1996: 7) This is followed by a question then asked about total income from remittances or payments back home that had been received by the household over the past year. In 2001, these sources of household income were included as part of the prompt for the personal income question. The meta-data file states: "Income from the sale of home-grown produce or home-brewed beer or cattle was also to be included". If any of these activities brought in income for the household as a whole rather than for a particular person, the enumerator was instructed to add the amount to the income of someone in the household. If the household had received remittances or payments from a person working or living elsewhere, the instruction was that this income should be added to the total of someone in the household, for example, the head of the household. (Statistics South Africa, 2000a: 81) Given these differences between 2001 and 1996, the personal income data is not directly comparable. Aggregate household income data, including the two household level questions in 1996, should theoretically be comparable. However, it is unlikely that without the specific questions around household level income, that such income was thoroughly captured in 2001. For this reason, we decided to use only income collected at the individual level in 1996. It should therefore be borne in mind that the 1996 estimates of per capita income are likely to be understated.

[5] Leibbrandt *et al* (2004) compressed the top end of the 2001 distribution of personal incomes into the real income equivalent of the top band in 1996. As all of these bands are way above any plausible poverty line, this has no impact on the analysis of poverty. However, as this decision effectively compresses the top end of the 2001 income distribution, this decision impacts on the inequality analysis. See Table A.3 in Appendix A of Leibbrandt *et al* (2004) for a detailed set of results.

The paper explores two further weaknesses in the personal income variable; namely, the large number of working age adults for whom the income variable is missing and the large number of working age adults for whom recorded income is zero. As shown later in the paper, a large percentage of individuals are recorded as having missing incomes or zero incomes. On aggregating these personal incomes into household incomes, this translates into a large number of households with missing total income values or zero total income values.

It is important that these two issues receive detailed attention. Regarding the missing data, who are these people and households? If they were not missing, where would they have fallen in the distribution of income and what impact would they have had on measured poverty and inequality? Regarding the zero incomes, even allowing for South Africa's low labour market participation rates and high unemployment rates, it is highly unlikely that all of these zero income households had no adult members earning any income. In analysing poverty and inequality, previous practice has been to ignore the zeros or to change them to some arbitrarily small number. The former practice is an arbitrary decision to effectively remove a group of households who currently make up the bottom of the distribution. As such, this decision sharply decreases measured poverty levels and also narrows inequality. The latter practice effectively accepts all recorded zeros as genuine zeros, possibly leading to an overestimate of measured poverty and inequality.[6]

In sum then, the focus of this paper is on three weaknesses in the personal income variable in the 1996 and 2001 South African census data; namely, missing data, a large number of implausible zero values and the fact that income is measured in bands. All of these weaknesses impact on measured individual and household income and therefore on measured poverty and inequality.

In the next section of the paper, we deal with missing data by imputing income bands for those with missing income data for 2001 using contemporary multiple imputation techniques. Statistics South Africa offers users of the 2001 data a single hotdeck imputation for the missing 2001 personal income data. In line with contemporary practice, multiple imputation approaches are preferred to single imputations. Our work in this section will discuss and use a multiple imputation approach and will compare our imputation results with the hot deck results of Statistics South Africa. In the third section of the paper we consider the impact of implausible income values; in particular, the high percentage of households with zero income. Our approach is to use a set of decision rules to reclassify potentially problematic zero incomes as missing and then to re-run the multiple imputation process on the augmented missing data. The process allows

[6] An example of the impact of these assumptions on measured poverty and inequality is contained in Appendix A of Leibbrandt *et al* (2004).

for the possibility that any values that are reclassified from zero to missing to be imputed back into the data as a zero income once more if the census data support such an imputation.

The income data in the 2001 census is given in twelve bands. As stated previously, in order to estimate measures of poverty and inequality, a continuous measure of income is required. Therefore, a further "imputation" step is required in order to translate the bands into point estimates. The lowest income "band" is zero income and no within-band decisions are necessary here. For the next ten bands, the convention (including for our analysis in sections 2 and 3 of this paper) is to allocate to each individual the midpoint income of the band within which the person is found. Finally, incomes falling in the highest (unbounded) band are all assigned the lower bound value for this top band. In section 4, we examine the sensitivity of the key results that are derived using this set of rules to those that are derived when we impute within-band point estimates from empirical distributions of personal incomes in each band. These empirical distributions are available from a national household income and expenditure survey of 30,000 households that was conducted in 2000.

One of the most important uses of census data is to calculate provincial poverty shares. As a final exercise on the 2001 data, in Section 5 we consider the impact of our imputations on the estimates of provincial poverty shares. These shares are important from a policy perspective as they are central to the formula for allocating budget allocations for anti-poverty programmes. Encouragingly, we find that provincial poverty shares are robust to a range of assumptions about missing data values and the distribution of incomes within bands.

In order to keep the discussion of sections 2 through 5 manageable, we discuss the techniques and illustrate their impact using the 2001 census data. However, all exercises were replicated on the 1996 data. The final section of the paper briefly returns to the issue of comparing 1996 and 2001 poverty and inequality situations in the light of our imputation work.

2. Dealing with Missing Data

The potential bias in estimates caused by missing data is a pervasive problem in empirical work. Unless the data is missing completely at random (MCAR), estimates that exclude individuals with missing data from the analysis will be biased.[7] Missing data is particularly problematic in calculating measures of

[7] Suppose that y_i is a response of interest (income in this case), x_i is a vector of information (province, rural/urban, race, age, sex, education, employment status, occupation) known about

poverty and inequality that are sensitive to the full distribution of the data. If those with missing data fall disproportionately in the bottom of the distribution, then levels of poverty will be underestimated. Alternatively, if non-response is higher among the wealthy, measures of inequality are likely be biased downwards. Sixteen percent of individuals in the 2001 census ten percent sample have missing income data. The missing data problem is exacerbated with analyses at the household level as more than a quarter of individuals belong to households where all or some of the household members have missing income data. If missing data is ignored, all these individuals are excluded from any household level analyses such as the calculation of per capita poverty and inequality measures. Table A2 in the Appendix shows the rates of non-response on the income question across various categories of individuals. There are significant differences in the response rates across a number of variables. Whites are much more likely to have missing income data (24%) than Black Africans (14%). Response rates are higher in rural areas than urban areas. There is large variation in response rates across provinces with more than 23% of individuals in the Western Cape missing income data and less than 7% of individuals in the North West. It is clearly evident from the missing data patterns in Table A2 that the income data is not MCAR.

If the data is missing at random (MAR), then we can adjust for non-response in order to reduce bias in our estimates[8]. There are a range of methods for handling missing data including weighting, imputation and non-parametric techniques (Lohr 1999, Little and Rubin 2000). The 2001 ten percent sample from Statistics South Africa offers a set of imputations for all the missing data. In this case, a single-imputation hot deck method was used to impute missing income values for individuals.[9] This means that missing values "are replaced by values from

person *i* in the sample. If the probability that person i will respond does not depend on x_i, y_i or the survey design, the data are MCAR. If data are MCAR, the respondents are representative of the selected sample. The MCAR mechanism is implicitly adopted when non-response is ignored (Lohr 1999: 264).

[8] If the probability of response depends on x_i, but not on y_i, the data are MAR as the non-response depends only on observed variables. We can successfully model the non-response, since we know that values of x_i for all sample units. If the probability of non-response depends on the value of y_i and cannot be completely explained by values of the x's, then the non-response is non-ignorable as we are unable to model it. (Lohr 1999: 265)

[9] Statistics South Africa used a combination of two kinds of imputations with the 2001 data. The first were "logical" imputations and the second were "hotdeck" imputations. For the logical imputations, "a consistent value is calculated or deduced from other information relating to the individual or household. For example, a married person with invalid sex would be assigned to the opposite sex of his or her spouse." (Statistics South Africa 2001a: 3) If a logical imputation was not possible then the "hotdeck" procedure was used. While the conceptual distinction between logical imputations and other imputations is clear, there is not sufficient documentation on the rules that Statistics South Africa used for their logical imputations. In addition, tabulations of the logical imputations for some of our variables

similar responding sampling units. The hot deck literally refers to the deck of matching computer cards for the donors available for a nonrespondent" (Little and Rubin, 2002: 66).

In this paper, we use a multivariate regression technique to multiply impute missing values. This technique has a number of advantages over the single hot deck imputation approach adopted by Statistics South Africa. Firstly, a multivariate multiple imputation approach is more robust than a single hot deck imputation. While all imputation based approaches rely on the observed data to impute values for missing items, the hotdeck technique is particularly sensitive to the problems of badly measured variables. If the data on some of your respondents is not good then there is a chance of you drawing a bad respondent as a donor to replace your non-respondent. For example in the 2001 data, most fifteen years olds are not earning any income. However, of the few that report that they are employed and are earning, one in two are earning implausibly high levels of income. With these values in the data set, a fifteen year old with a missing income value might draw one of these cards from the hot deck and be given an implausibly high income. In this way, hot deck imputation might be magnifying the problems of badly measured variables. For each multivariate regression imputation, the impact of these outliers in generating imputed values is lessened. In addition, final estimates are obtained by averaging over multiple imputations, further reducing the problems of badly measured variables.

Secondly, any single imputation technique does not distinguish between observed and imputed values in the resultant data set and as such the variance of any estimates is understated. Multiple imputations generate a distribution of imputed values and a distribution of parameters of interest. This allows for a measure of the uncertainty due to imputation to be reflected in the standard errors of the estimates. Given such advantages, the imputation literature has a strong preference for running multiple imputations using a suitable multivariate technique (Little and Rubin 2002).

2.1 The Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models

The multiple imputation approach adopted in this paper follows the sequential regression multiple imputation (SRMI) approach of Raghunathan *et al* (2001).

throw up some results that are not immediately obvious. For example, it is clearly "logical" to code two year olds as having no education but it is not clear how a number of adults were "logically" imputed to have no education. Therefore in this paper, we do not distinguish between Statistics South Africa's "logical" and "hotdeck" imputations.

Most directly, our task is to impute an appropriate personal income band for each individual with missing income data. As mentioned above, in the 2001 census, there are 12 income bands with 0 being the lowest "band" and R204 801 a month (2 457 601 a year) being the lower bound of the top band. Thus, this task is to specify an ordered logit model that uses the best set of variables that are available in the census to allocate missing data into these income bands.[10] The explanatory variables used in the ordered logit can be broken into two sets; an X and a Y matrix. The X matrix contains the set of predictor variables with no missing values. The Y matrix contains the variables with missing values. Let the *k* variables $Y_1, Y_2, \ldots, Y_k$ represent these variables ordered by the amount of missing values from least to most.

With the census data we have complete data for each person on province of residence, whether they resided in an urban or rural area and race.[11] These variables therefore make up the X matrix. The set of Ys ordered from least to most missing values consisted of age (a count variable), a gender dummy variable, an employment dummy variable[12], a four category occupation variable[13], years of education[14] (a count variable), and income (an ordered

[10] Van Buren *et al* (1999: 5) summarise a literature showing that "including as many predictor variables as possible tends to make the MAR assumption more plausible". Given that most data sets are very large, computationally, it is not really possible to include every possible variable. However it is also not necessary as the increase in explained variance is often minimal once the best set of variables have been included. Therefore, at the very least, one wants to include the best set of variables that are available in the census for predicting income bands for individuals. We include 9 variables which is a little below the maximum of 15 to 25 variables that van Buren *et al* (1999) suggest as a rule of thumb.

[11] Although one percent of individuals were missing values on the race variable, we chose to use Statistics South Africa's imputations. The majority of the imputations were "logical" and in most instances individuals were assigned the race of other people in their household. Our decision to use Statistics South Africa's imputations for race was practically motivated. Given the multiple imputation process, imputing missing values for race would require the fitting of 25 multinomial logit models, each with 26 independent variables (8 dummies for the provinces, 11 dummies for income bands, 3 dummies for occupation, education, age, sex and location) for each imputation. With a data set as large as the 10 percent micro-sample, the computational requirements are very demanding. As our interest was in imputing and analysing income and not race, we felt that it would be acceptable to use Statistics South Africa's imputations.

[12] This is a dummy variable derived from question P-18 in the census that codes people who did "any work for pay (in cash or in kind), profit or family gain for one hours or more" in the seven days preceding the Census as 1 and all others as 0. Statistics South Africa (2001a, p. 49) Therefore the zero includes unemployed and non-participants. As the questions on work were only asked for persons aged 10 years or older, all individuals under 10 years of age were coded by us as not employed.

[13] Following question 19 in the census, occupation equals 1, for legislators, senior officials and managers and professionals, 2, for elementary occupations, 3, for all other occupations and 4 for people with no occupation.

categorical variable of 12 income bands). The reason for the inclusion of income in the Y matrix goes to the heart of the Rangunathan *et al* (2001) SRMI approach to imputation. In this approach, income is imputed as part of a process of imputing missing values for all of the variables in the Y matrix. All missing values are imputed as part of a process to estimate the joint conditional density of Y_1, Y_2,…,Y_k given X. This density can be factored as:

$$f(Y_1,Y_2,\ldots,Y_k|X,\beta_1,\beta_2,\ldots,\beta_k)=f_1(Y_1|X,\beta_1)f_2(Y_2,|X,Y_1,\beta_2)\ldots f_k(Y_k|X, Y_1, Y_2,\ldots,Y_{k-1},\beta_k) \qquad (1)$$

where f_i represent the conditional density functions and β_i is a vector of parameters in the conditional distribution. In all cases the β_i vectors are estimated coefficients as well as estimates of the disturbance term. As mentioned above, our Y matrix contains count variables (age and education), binary categorical or dummy variables (gender, employment), a multiple category variable (occupation) and an ordered categorical variable (income). If Y_i is a count variable, then a poisson distribution is used to estimate f_i. If Y_i is binary, f_i is estimated using a logistic distribution. If Y_i is categorical, a multinomial logistic regression model is estimated and if Y_i is ordinal, an ordinal logistic regression model is estimated.

Settling on a set of imputed values for the missing Ys is analogous to settling on a satisfactory estimate of the joint conditional density of Y given X. The model is settled over a number of rounds. As reflected in (1) above, the first round starts with obtaining an estimate of the vector β_1 in a regression of Y_1 on X. The missing values in Y_1 are then replaced by random draws from the posterior predictive distribution. That is, by first drawing a vector β_1^* from the posterior distribution of β_1 and then using β_1^* to generate a set of predicted values to replace the missing Y_1 values. This is followed by an estimate of Y_2 given X and the newly derived Y_1, including imputed values, on the right hand side. The first round finishes when the ordered logit is estimated to derive initial imputed values for missing Y_k (income bands) conditional on X and Y_1 to Y_{k-1}. Once this model has been estimated, the first complete data set with no missing values is available.

At the start of the second round, Y_1 is re-estimated including all first round Y imputations on the right hand side. The first round missing value imputations for Y_1 are replaced by a new set of imputations derived from this re-estimation. All in all, the essence of the Ranghunathan *et al* (2001: 88) method is that "the new imputed values for a variable are conditional on the previously imputed values

[14] Education questions were only asked of people aged six or older as this covers the usual starting age for school in South Africa. Therefore, individuals aged 5 and under who were missing education data, were coded as having zero years of education in advance of the imputation process.

of the other variables and the newly imputed values of variables that preceded the currently imputed variable". These authors state that although it is theoretically possible that such a process does not converge to a stationary distribution, they have never encountered this in an empirical setting. As recommended by Raghunathan *et al* (2001), we ran five rounds or iterations before settling on final imputed values.

This is merely the start of the multiple imputation process. This is the sequential regression equivalent to the single vector of hot deck imputations that were derived by Statistics South Africa. We need to generate a set of such imputations to give us a distribution of imputed values and a distribution of parameters of interest. Ranghunathan *et al* (2001) recommend four or five imputations and we derived a set of five imputed data sets. Each imputation starts in the same way as described above by estimating a starting value for Y_1. Clearly the first estimated regression coefficients will always be the same but as the missing values are imputed after each regression using a random draw from the posterior distribution of the regression coefficients, a different set of imputed values is generated. Thus, from the very first imputation, this round generates different imputed values from the first round.

Thus, at the end of this process, five complete data sets were derived that incorporated five equally plausible sets of imputations for income bands for all individuals who had missing values on the income variable. Parameters of interest can be derived from each data set. In the context of this paper, three parameters of interest are mean household per capita income, an index of poverty (as measured by the head count index) and an index of inequality (as measured by the Gini coefficient). The multiple imputation variance formula (see Little and Rubin (2002: 86)) suggests that, in each of these three cases, the best estimate of each multiply imputed parameter is the mean of the five estimates of that parameter. The variability associated with this estimate consists of two parts; namely, the average within-imputation variance (Vw) and the between-imputation variance (Vb). A within sample variance is calculated for each parameter of interest each of the five times the parameter is calculated. The Vw is the mean of these five variances. The Vb is calculated as the variance of the five parameter estimates. The total variance equals Vw + ((m+1)/m)Vb where m is the number of imputations and ((m+1)/m) is an adjustment for the fact that the Vb is being calculated off a finite number of parameter estimates. The square root of this total variance is the standard error associated with the best estimate of each of our three parameters. In any single imputation model, one would be offered a single parameter estimate and a single variance associated with this estimate with no sense of how this parameter estimate might vary across equally plausible sets of missing data imputations.

2.2 Estimates of Mean Per Capita Income, Poverty and Inequality

The technique outlined above was used to derive a set of imputed values and ascertain their influence on measures of poverty and inequality. Table 1 presents point estimates and 95% confidence intervals for a range of poverty and inequality measures for 2001. Mean per capita income and Gini coefficients are presented in Panel A and poverty headcounts for both a R124 and R400 poverty line are shown in Panel B. The first row presents estimates that were calculated ignoring all missing income data. The estimates in the second row were calculated using Statistics South Africa's hot deck imputations. The third row presents the combined estimates from our multiple imputations. The estimates for each of the five imputed data sets are also shown.

The imputations suggest that ignoring the missing values results in downwardly biased estimates of mean income and inequality and upwardly biased estimates of poverty at both poverty lines.[15] When we adjust for non-response, mean per capita income increases, the percentage of households in poverty decreases and inequality decreases. These results are consistent with our previous observation that response rates were lower in urban areas and amongst Whites, suggesting that individuals with imputed incomes have on average higher incomes than individuals with non-missing income.

The parameter estimates for Statistics South Africa's hotdeck imputation are closer to the combined SRMI estimates than to the no imputation results. However, the confidence intervals for the hot deck estimates are noticeably tighter than those of the combined multiple imputation estimates. This is because the variances for the hotdeck results are akin to the within-imputation variances of any of the five imputations that are derived as part of the SRMI multiple imputation process. The between-imputation variance is ignored by the single imputation hotdeck. This variance is clearly seen by comparing the estimates across each of the multiple imputations. Clearly, the single imputation technique that does not take into account uncertainty due to imputation overstates the precision of the estimates. Multiple imputation techniques take into account both the potential bias and the uncertainty due to missing values. The fact that the results on the mean, the head count ratio and the Gini

[15] Results for the poverty gap ratio are presented in Table A3 of the appendix. The poverty gap ratio is sensitive to the gap between the incomes of the poor and the poverty line. As such, is it sensitive to the value of each imputed income rather than merely whether such income is above or below the poverty line and it could be more sensitive than the headcount ratio to our imputations (see Foster *et al* (1984)). However, our poverty results are the same when redone using the poverty gap ratio rather than the headcount ratio.

coefficient remain intact even allowing for uncertainty due to multiple imputations is very helpful in defending these estimates.

Table 1 Comparison of poverty and inequality measures 2001

PANEL A. Mean per capita income and Gini coefficients						
	Mean per capita income			*Gini Coefficient*		
	Estimate	*95% C.I.*		*Estimate*	*95% C.I.*	
No imputed values	910.457	909.041	911.873	0.773	0.772	0.774
Statistics South Africa's hot deck imputation	1033.606	1032.325	1034.886	0.819	0.818	0.819
SRMI Multiple imputation						
Combined	*1015.074*	*926.548*	*1088.848*	*0.819*	*0.815*	*0.822*
Imputation 1	1084.292	1082.924	1085.660	0.822	0.821	0.823
Imputation 2	1002.964	1001.712	1004.216	0.818	0.817	0.818
Imputation 3	973.442	972.213	974.670	0.818	0.817	0.818
Imputation 4	1005.381	1004.122	1006.640	0.818	0.817	0.819
Imputation 5	1009.289	1008.026	1010.552	0.818	0.817	0.819

PANEL B: Poverty Headcount Ratios						
	Poverty Headcount Ratio (Poverty line at R124 per capita per month)			*Poverty Headcount Ratio (Poverty line at R400 per capita per month)*		
	Estimate	*95% C.I.*		*Estimate*	*95% C.I.*	
No imputed values	0.452	0.451	0.453	0.685	0.685	0.686
Statistics South Africa's hot deck imputation	0.422	0.421	0.422	0.656	0.656	0.657
SRMI Multiple imputation						
Combined	*0.425*	*0.413*	*0.437*	*0.659*	*0.648*	*0.669*
Imputation 1	0.416	0.416	0.417	0.651	0.650	0.651
Imputation 2	0.426	0.426	0.427	0.659	0.659	0.660
Imputation 3	0.432	0.431	0.432	0.664	0.664	0.665
Imputation 4	0.426	0.426	0.427	0.660	0.659	0.660
Imputation 5	0.426	0.425	0.426	0.659	0.658	0.659

Source: Census 2001 (authors' own calculations).

Note: A continuous measure of personal income was generated by allocating each individual the midpoint income of the band within which they are found. The highest (unbounded) band was assigned the lower bound value. Furthermore, because we are interested in per capita income, we summed all positive personal income for each household and then divided by household size to obtain a monthly per capita measure of income. For comparability between the two censuses and to avoid problems in calculating household size, we excluded all data on people living in institutions, and all results were weighted using the weights supplied by Statistics South Africa.

3. Assessing the Importance of Implausible Values

Given our warnings above about imputing off a data set that contains outliers and implausible values, as a next step we investigate the sensitivity of our results to such values in the data set. This is an especially important stage in the analysis of income as it allows us to acknowledge and deal with the implausibly high proportion of zero income households that are recorded. These households clearly have an impact on estimates calculated from the observed data values. Our imputations above (and all imputation approaches) rely on the observed data and based on this data, it is likely that a large number of households with missing income data will be imputed to have zero income. In this section, we present a fresh set of imputations that begins by taking problematic zero values and recoding them as missing before any imputation takes place. In this way, zero income households are screened for plausibility and then either are assigned a positive income amount or affirmed as a zero income household through the imputation process.

Clearly there are some households that genuinely earn zero income and by setting all individuals in such households to missing, we remove these valid observations from our observed data, thus affecting our imputations. These observations only come back into the imputation process at the end of the first round of regression estimations and there is some chance that they are imputed to have positive incomes at this point. In other words, our screen for plausibility has some biases against the zero income households. In this sense the imputations generated in this section cannot be seen as deriving an unambiguously superior set of estimates, but rather as investigating the sensitivity of our results to outliers and implausible values in the observed data.

We began by recoding problematic values to missing using the following rules:

- If household income was zero, income was set to missing for household members aged 15 and older and to zero for those younger than 15.
- For those younger than 15 with recorded income greater than R6 400 per month, income was set to missing.
- For those recorded as being employed but with zero income, we set income to missing.

This gave us a new base data set of individuals in which income was coded as missing or as one of 12 income bands. We then undertook the same multiple imputation process as before on this new base data set in order to transform all missing values into one of the 12 income bands.

Table 2 presents the percentage of households and individuals whose income values warranted closer inspection. The table shows the percentage and number of households reporting zero income, the percentage of employed people earning zero income and the percentage of people aged 15 and under earning over R6 400 per month. In all cases, these are national figures derived using sampling weights. These percentages were calculated under four different data assumptions. The first row presents estimates based on a complete case analysis where the missing data was ignored. The second set of estimates use Statistics South Africa's hotdeck imputations. The third row presents the combined estimate of our five imputed data sets of the previous section of the paper. The final row presents the combined estimate for our new set of imputations where implausible values were reset to missing as described above.

For both the hotdeck imputation and our multiple imputation of the previous section, the percentage in each category is very similar to the data set where missing values were ignored. The dependence of the imputed values on the observed data is clear. In the second multiple imputation where implausible values were set to zero, the proportion of households or individuals in each category is significantly reduced. While the imputation process allows for some households to be reclassified as earning zero, many of the households previously classified as earning zero income are now imputed to earn some positive income. As stated above, this should be viewed as a lower bound and we would expect the true percentage of zero income household to lie somewhere between 14% and 23%. The results for employed people earning zero income and high income children are similar.

The analysis of high income children shows very interesting differences across each category. With no imputations, 0.14% of the population aged 15 and under is captured as such high earners. Earlier in the paper, it was mentioned that one in two fifteen year olds that reported positive incomes were earning implausibly high amounts. It was mentioned that this situation resulted in the hotdeck imputation having a high probability of drawing high earning fifteen year olds in its imputation of missing values. Given this possibility, it is interesting to note that the percentage of high earning children increases to 0.17% (about 9 000 children) as a result of the hotdeck imputation. In the first multiple imputation 1 process, about 1 000 additional children are imputed to be high-earning. However, this group now represents a smaller percentage of the population aged 15 and under (0.12%). In the second multiple imputation, these high-earning children are set to missing at the start. Given this new base data set, the multiple imputation process gives high earnings to a mere 0.02% of children. Clearly, this is a lower bound value rather than a clearly more correct value. Nonetheless, the well-being of children is a key issue in South Africa and this demonstration of the sensitivity of estimates of children's income to a high-earning group in the data set is an important cautionary note.

Table 2 Percentage and number of households reporting zero income, employed people earning zero income and people aged 15 and under earning in excess of R6,400 per month

	Households reporting zero income	*Employed people earning zero income*	*People aged 15 and under with incomes in excess of R6 400 per month*
No imputed values	2 168 820 (25.27%)	157 834 (1.90%)	16 493 (0.14%)
Hotdeck imputation	2 541 034 (23.18%)	231 560 (2.39%)	25 510 (0.17%)
SRMI Multiple imputation	2 553 678 (23.30%)	209 359 (2.28%)	17 707 (0.12%)
SRMI Multiple imputation with implausible values set to missing	1 540 786 (14.06%)	52 307 (0.57%)	2 101 (0.02%)

Source: Census 2001 (authors' own calculations).
Note: All results were weighted using weights supplied by Statistics South Africa.

Table 3 contrasts measures of poverty and inequality for each set of multiple imputations. Combined multiple imputation estimates calculated in the previous section are shown in the first row of each panel of the table. The second row presents the combined estimates from five imputations where implausible values were set to missing at the start of the imputation process. The results for each of the five imputations are shown below the combined estimates. The impact of recoding implausible values to missing is seen in an increase in mean per capita income and a decrease in both the Gini coefficient and the percentage of households in poverty. It is interesting to note the increases in mean income given our earlier discussion about the greatly reduced number of high-earning children in the second imputation. However, one should not make too much of these point comparisons as, only the difference in Gini coefficients is statistically significant.

Perhaps the key point to note from this exercise is that the precision of the estimates for the second set of imputations is much reduced. Inspection of the set of multiple imputations in Table 3 shows that the increase in the combined mean per capita income is driven by one unusually large estimated per capita income in imputation 4. It is a strength of the multiple imputation approach that this uncertainty introduced into the combined estimate of mean per capita income drives up the variance of the combined estimate and is reflected in Table 3 by the large confidence interval on the combined mean per capita income estimate. Applying the rules set out above increases the rate of missing income data to 35% of individuals. In general, the higher the rate of missing data, the greater between-imputation variance. The increased uncertainty due to a higher rate of missing data substantially reduces the precision of the estimates. This clearly illustrates the point that the use of missing value imputation techniques to assess the influence of non-missing but implausible values in the data has to be very well motivated and, even then, must be regarded as a sensitivity exercise rather than the production of better estimates. In this case, the results in Table 3 are encouraging in that our sensitivity analysis reveals that the implausible values do not have a significant impact on the substantive interpretation of our estimates of interest.

Table 3 Comparison of poverty and inequality measures 2001

PANEL A. Mean per capita income and Gini coefficients						
	Mean per capita income			*Gini Coefficient*		
	Estimate	*95% C.I.*		*Estimate*	*95% C.I.*	
SRMI Multiple imputation	1015.074	926.548	1088.848	0.819	0.815	0.822
SRMI Multiple imputation (implausible values set to missing)						
Combined	*1056.524*	*749.108*	*1312.706*	*0.800*	*0.793*	*0.805*
Imputation 1	1034.586	1033.334	1035.838	0.797	0.796	0.797
Imputation 2	964.476	963.287	965.664	0.803	0.803	0.804
Imputation 3	997.360	996.148	998.572	0.798	0.797	0.799
Imputation 4	1308.258	1306.639	1309.877	0.798	0.798	0.799
Imputation 5	977.942	976.741	979.142	0.803	0.802	0.804

PANEL B. Poverty Headcount Ratios						
	Poverty Headcount Ratio (Poverty line at R124 per capita per month)			*Poverty Headcount Ratio (Poverty line at R400 per capita per month)*		
	Estimate	*95% C.I.*		*Estimate*	*95% C.I.*	
SRMI Multiple imputation	0.425	0.413	0.437	0.659	0.648	0.669
SRMI Multiple imputation (implausible values set to missing)						
Combined	*0.375*	*0.279*	*0.471*	*0.640*	*0.588*	*0.691*
Imputation 1	0.376	0.375	0.376	0.641	0.641	0.642
Imputation 2	0.408	0.408	0.409	0.657	0.656	0.657
Imputation 3	0.388	0.388	0.389	0.648	0.648	0.649
Imputation 4	0.298	0.298	0.299	0.598	0.598	0.599
Imputation 5	0.405	0.404	0.405	0.654	0.654	0.655

Source: Census 2001 (authors' own calculations).
Note: All results were weighted using weights supplied by Statistics South Africa.

4. From Bands to Point Incomes

In both the 1996 and 2001 census, income was collected in bands. In order to estimate measures of poverty and inequality, a continuous measure of income is required. Therefore, a further "imputation" step is required in order to translate the bands into point estimates. The conventional approach adopted by most applied researchers is to assign individuals the midpoint income of the band within which they are found. Alternatively, individuals can be assigned incomes within their band according to some other intra-band distributional rule. In this section, we examine the sensitivity of measures of poverty and inequality to different assumptions about the distribution of income within the bands.

In sections 2 and 3 of this paper we generate a continuous measure of personal income by allocating each individual the midpoint income of the band within which they are found. The lowest band is zero income and is therefore already a point income. The highest (unbounded) band was assigned the lower bound value for this band. All in all these rules represent the conventional approach to these bands.

An alternative intra-band allocation rule requires the generation of point incomes by randomly sampling from a specified distribution within each band. While the uniform, normal or lognormal distributions are often used, there is generally no dominant theoretical basis for the choice of distribution or the value of its parameters. Indeed there is no theoretical basis for the conventionally adopted mid-point approach. Ideally one would want the distribution to be as close as possible to the true distribution of personal incomes. We therefore decided to use an empirical distribution based on another appropriate data set; the 2000 Income and Expenditure Survey (IES)[16]. Personal incomes in the 2000 IES were first adjusted to 2001 equivalents using a single price inflator.[17] Then, an empirical cumulative distribution was generated for each income band. Random probabilities were generated for each individual in the Census 2001 data and individuals were assigned an income such that the cumulative probability of observing such a value from the empirical distribution was greater than or equal to the generated probability. It is important to note that this does not replicate the full distribution of personal incomes within the IES but rather

[16] Total regular personal income was used to generate an empirical distribution (Statistics South Africa 2000 Section 24.1: 44-47). While this excluded household level income, it is not clear that household level income would have been well captured by the 2001 Census (see footnote 3). Importantly, we were not trying to impose the empirical distribution of personal income in the IES onto individuals in the Census but rather to match the IES empirical distribution **within** each income band.

[17] The percentage chance in CPI between October 2000 and October 2001 was 4% (Statistics South Africa 2001b: 1)

replicates the intra-band IES distribution. This approach would seem to be particularly useful in imputing point incomes for those in the top band. While decisions about this top band have no implications for poverty, measured inequality is likely to be sensitive to changes at the top of the income distribution.

Table 4 Comparison of poverty and inequality measures 2001

PANEL A. Mean per capita income and Gini coefficients						
	Mean per capita income			*Gini Coefficient*		
	Estimate	*95% C.I.*		*Estimate*	*95% C.I.*	
SRMI Multiple imputation (mid-points)	1015.074	926.548	1088.848	0.819	0.815	0.822
SRMI Multiple imputation (IES empirical distribution)						
Combined	*1002.000*	*931.036*	*1061.146*	*0.822*	*0.817*	*0.827*
Imputation 1	1038.799	1037.157	1040.441	0.826	0.825	0.827
Imputation 2	1012.155	1009.999	1014.311	0.821	0.819	0.822
Imputation 3	981.567	979.599	983.536	0.820	0.819	0.822
Imputation 4	956.029	954.546	957.512	0.821	0.820	0.822
Imputation 5	1021.448	1019.213	1023.683	0.822	0.821	0.823

PANEL B. Poverty Headcount Ratios						
	Poverty Headcount Ratio (Poverty line at R124 per capita per month)			*Poverty Headcount Ratio (Poverty line at R400 per capita per month)*		
	Estimate	*95% C.I.*		*Estimate*	*95% C.I.*	
SRMI Multiple imputation (mid-points)	0.425	0.413	0.437	0.659	0.648	0.669
SRMI Multiple imputation (IES empirical distribution)						
Combined	*0.417*	*0.398*	*0.437*	*0.675*	*0.663*	*0.687*
Imputation 1	0.421	0.421	0.422	0.675	0.674	0.675
Imputation 2	0.410	0.409	0.410	0.670	0.670	0.671
Imputation 3	0.415	0.415	0.416	0.675	0.675	0.676
Imputation 4	0.431	0.431	0.432	0.684	0.683	0.684
Imputation 5	0.409	0.409	0.410	0.670	0.669	0.670

Source: Census 2001 (authors' own calculations).
Note: All results were weighted using weights supplied by Statistics South Africa.

Theoretically, we cannot predict whether assigning all imputed incomes to the midpoints of bands will generate more or less inequality than distributing the incomes across the bands. While the effect at the top band is clear – assigning everyone to the bottom cutoff of the band must generate less inequality than distributing individuals across the band, the effect in lower bands is indeterminate. While individuals within a band get compressed when midpoints are used, many individuals in adjacent bands will be spread further apart when midpoints are used. The net effect is theoretically ambiguous. Similarly, we cannot predict *a priori* which of the two approaches will generate the higher poverty headcount. This will depend on the position of the poverty line relative to the midpoint of a band, the movement across that threshold caused by adjustments for household size, and other complex interactions between the poverty line and the income imputations. It is worth noting that the R400 poverty line corresponds to the top of the second income band (which goes from R1 to R400), while the R124 poverty line is below the midpoint of this band. While the effects of using midpoints versus full distribution across the band would be relatively easy to analyse in the case of individual incomes and an individual poverty line, it is considerably more complicated in the case where we are using aggregate household income adjusted for household size.

To assess the influence of these allocation rules, we use the five data sets that were generated in multiply imputing all missing income values in section 2 of the paper. As the midpoint rule was used for intra-band point income allocation in section 2, Table 4 simply reproduces the key parameter results from this section to represent this mid-point case. Table 4 also presents the resultant parameter estimates obtained using the empirical distribution of personal income in the IES within each band on each of the five multiply imputed data sets. The combined parameter estimates and confidence intervals are derived from these five estimates in the same way that they have been for all of the multiple imputations in this paper. Estimates of poverty gap ratios are presented in Table A3 in the Appendix.

Comparing the two sets of combined estimates it can be seen that the point estimate of mean per capita income is lower using the IES distribution. The falling mean suggests that on average within bands, the distribution of income is skewed to the right with the mean below the mid-point of the band. However, one should not make too much of these point comparisons as, while these point estimates are lower using the IES distribution rule, the differences are not statistically significant. The use of the IES distribution does not produce uniform or significant changes in the poverty headcount or poverty gap ratio.

When the IES rule is used, the estimated Gini coefficient rises; suggesting an increase in inequality. As noted above, this will in part be attributed to the fact that the IES rule stretches out personal incomes at the top end, with effects in other bands either reinforcing or offsetting the increase in inequality at the top. Again, not too much can be made of this increase as the difference between the two combined Gini coefficients is not statistically significant.

Thus, in sum, it does not appear that our results are very sensitive to the assumptions underlying the "imputation" of a continuous variable from the bands. However, as the distribution of income within bands is more likely to follow the empirical distribution of personal income in the IES, we prefer this technique to the mid-point approach.

5. Provincial Poverty Shares

In sections 2 through 4 we have imputed missing values, checked for the sensitivity of results to dubious zero values by setting these to zero and re-running the imputations and then assessed the sensitivity of key parameters estimates to two intra-band point allocation rules. The results suggested that, at the national level, all imputations significantly increased the estimates of mean income and the Gini coefficient measure of inequality and decreased measured poverty. Generally, the results are not statistically different from each other across these imputation exercises and certainly do not show changes that are large enough to be socially significant.

In the practice of contemporary development policy in South Africa, there is one specific policy exercise that makes intensive use of the 2001 census data. Provincial and municipal poverty shares feature in the anti-poverty budget allocation formula that are utilised by the national treasury to fund poverty alleviation programmes. Thus, as a final exercise in this paper, it seems appropriate to assess the sensitivity of provincial poverty shares to our imputations. There is another good reason to do this. Table A2 in the Appendix describes the extent of missing income values in the census data. We used this table to argue that the missing data were not missing completely at random as non-response differed across a range of variables, including province. Therefore, the assessment of the sensitivity of provincial poverty shares would seem to offer a policy relevant assessment of these influences as one begins to move from the national level to the disaggregated level.

Table 5 presents four sets of combined multiple imputation estimates of provincial poverty measures and shares for 2001. Each was calculated using two poverty measures; namely, the headcount ratio and the poverty gap ratio. Foster *et al* (1984) describe how a provincial poverty share can be calculated as the product of a provincial poverty incidence measure (head-count or poverty gap) weighted by that province's share of the national population and then divided by the relevant national poverty incidence measure. As both the provincial poverty measure and the national poverty measure have their own standard errors, it is not possible to calculate standard errors and confidence intervals for these shares. However, as national budget rules do not give account to standard errors, this is not too unfair. Table 6 shows the results for a R400 per capita per month poverty line.[18]

Table 5 Provincial Poverty Shares 2001 with a Poverty Line of R400

PANEL A. Poverty Measured Using the Headcount Ratio								
	No Imputation		*SRMI Multiple Imputation*					
			Mid points		*IES 2000 empirical distribution*		*Implausible Zeros set to missing*	
Province	*Ratio*	*Shares*	*Ratio*	*Shares*	*Ratio*	*Shares*	*Ratio*	*Shares*
Western Cape	0.415	0.05	0.408	0.061	0.434	0.063	0.388	0.06
Eastern Cape	0.822	0.167	0.805	0.178	0.816	0.176	0.783	0.178
Northern Cape	0.652	0.018	0.636	0.018	0.655	0.018	0.621	0.018
Free State	0.757	0.069	0.734	0.068	0.75	0.068	0.719	0.069
KwaZulu-Natal	0.741	0.232	0.727	0.234	0.739	0.233	0.708	0.235
North West	0.71	0.097	0.698	0.086	0.717	0.086	0.684	0.087
Gauteng	0.473	0.126	0.443	0.132	0.464	0.135	0.42	0.129
Mpumalanga	0.745	0.081	0.728	0.076	0.745	0.075	0.714	0.076
Limpopo	0.835	0.161	0.821	0.147	0.832	0.145	0.806	0.149

PANEL B. Poverty Measured Using the Poverty Gap Ratio								
	No Imputation		*SRMI Multiple Imputation*					
			Mid points		*IES 2000 empirical distribution*		*Implausible Zeros set to missing*	
Province	*Ratio*	*Shares*	*Ratio*	*Shares*	*Ratio*	*Shares*	*Ratio*	*Shares*
Western Cape	0.256	0.041	0.245	0.05	0.251	0.05	0.216	0.048
Eastern Cape	0.64	0.173	0.621	0.185	0.625	0.185	0.564	0.185
Northern Cape	0.435	0.016	0.419	0.016	0.425	0.016	0.388	0.016
Free State	0.56	0.068	0.539	0.068	0.543	0.067	0.503	0.069
KwaZulu-Natal	0.576	0.241	0.559	0.244	0.563	0.243	0.511	0.244
North West	0.519	0.094	0.508	0.085	0.513	0.085	0.47	0.086
Gauteng	0.34	0.12	0.31	0.125	0.315	0.126	0.269	0.119
Mpumalanga	0.553	0.08	0.535	0.075	0.539	0.075	0.499	0.077
Limpopo	0.648	0.166	0.633	0.153	0.637	0.153	0.585	0.156

Source: Census 2001 (authors' own calculations).
Note: All results were weighted using weights supplied by Statistics South Africa.

The four major columns in the table represent, respectively, the no imputation calculations, the combined results from the multiple imputation of missing values where midpoints were used to calculate intra-band point incomes, the combined results from the multiple imputation of missing values where the IES

[18] We do not show the results for the lower, R124, line as these results are not sensitive to choice of poverty line.

intra-band allocation rule was used to calculate point incomes, and, finally, the results of the multiple imputation of missing values where these missing values had been augmented with dubious zero values.

Across these four, the major differences in the estimated poverty parameters are between the set of results derived without imputing missing values and the results from the three sets of imputations. In every province, the incidence of poverty is higher in the no imputation work than in any of the three imputations. This is true for either the headcount measure or the poverty gap measure. Similarly, the major differences in the poverty shares are not between the different imputation methods but between them and the no imputation shares. Provincial poverty shares represent a zero sum situation. Therefore, the fact that the no imputation results reflect larger shares for Limpopo, Mpumalanga and Gauteng implies that it has to deliver smaller shares somewhere else in the system. In this case, the lower shares are for Western Cape, Eastern Cape and KwaZulu-Natal.

The differences between the imputed versus the non-imputed poverty results are important as, to date, the convention has been to ignore missing values and to use no-imputation results. It is hard to make a case for this being a better practice than dealing with the missing values and, here we have showed that the treatment of missing data does impact on shares. This is not to say that the treatment of missing values is the only important issue. Indeed, Table 6 shows that the choice of poverty measure is another important source of variance in the provincial shares.

There are a few interesting points to note within the three different imputation processes. In each instance measured poverty is highest in the IES based imputations for missing data, followed by the midpoint based imputations for missing values, with the lowest estimated poverty parameters being found in the imputations that treated dubious zero income values as missing. The comparison of the midpoint and IES imputations suggests that, for those with household per capita incomes below the poverty line, the empirical distribution is imputing point values to a high proportion of persons in each band that are below the median value within each band. The comparison of these two imputations with the imputation that includes some dubious zeros suggests that a number of the dubious zero incomes are imputed to have positive income values even in poor households.

6. Discussion and Conclusion

We began this paper by summarising the literature that has assessed medium-run changes in poverty and inequality in South Africa using census data. According to this literature, over the 1996 to 2001 period both poverty and inequality increased. We pointed out that while this literature had paid considerable attention to issues of comparability over time, it had made little attempt to deal with the large percentage of individuals and households for whom personal income data was missing. We went on to use the sequential regression multiple imputation approach to impute missing values for the 2001 census data. The results suggested that, at the national level, the imputation increased the estimates of mean income and the Gini coefficient measure of inequality and decreased measured poverty. This was true even accounting for the wider confidence intervals that arise from the uncertainty that the imputations bring into the estimation process. There are a number of implausible income values in the data set. In the next section, we assessed the influence of these values by setting them to zero and re-doing the multiple imputation process with this augmented set of missing values. The missing values now made up 35% of all personal income observations. This clearly increased the uncertainty associated with the imputation process as reflected in wider confidence intervals on all estimates. As a result, the Gini coefficient was the only parameter of interest that was significantly different from the combined estimates obtained from the first set of multiple imputations.

Up to this point in the paper (and in the other relevant literature as well), all calculations had taken personal incomes recorded in bands and attributed band midpoints to all individuals. The final imputation exercise assessed the sensitivity of results to this implicit practice by using the empirical distributions of personal incomes that were available through a national income and expenditure survey and imputing intra-band point incomes in a manner that would replicate the intra-band distribution of income in this empirical distribution. Assigning midpoints of bands for imputed incomes leads to lower estimates of inequality than the alternative approach of distributing incomes across the entire band based on the empirical distribution from the IES, although the differences are relatively small. We finished our empirical work with a discussion of provincial shares as a policy relevant illustration of the importance of dealing with missing values.

The major goal of the paper was to deal with problems arising in using the personal income variable of the census due to the presence of a large amount of missing data. We introduced the sequential multiple imputation regression technique as an example of best international practice. In order to keep the discussion of the imputation technique manageable, in the main body of the paper we limited all empirical work to results using the 2001 data. However, all of the analysis recorded in this paper for 2001 was also undertaken on the 1996 census. The paper started out with reference to the work comparing poverty and inequality changes over time and it seemed appropriate to close out the paper with a brief comparison of multiply imputed poverty and inequality results for 1996 and 2001.

Table 6 Comparison of poverty in 1996 and 2001

	Poverty Headcount Poverty line: R124 in 2001, R91 in 1996			*Poverty Headcount Poverty line: R400 in 2001, R250 in 1996*		
	Estimate	*95% C.I.*		*Estimate*	*95% C.I.*	
SRMI Multiple imputation 1996 (IES 1995 empirical distribution)	0.383	0.382	0.383	0.600	0.599	0.602
SRMI Multiple imputation 2001 (IES 2000 empirical distribution)	0.417	0.398	0.434	0.675	0.663	0.685

Source: Census 1996; Census 2001 (authors' own calculations).
Note: All results were weighted using weights supplied by Statistics South Africa.

Table 7 Comparison of Inequality in 1996 and 2001

	Gini Coefficient		
	Estimate	*95% C.I.*	
SRMI Multiple imputation 1996 (IES 1995 empirical distribution)	0.744	0.743	0.745
SRMI Multiple imputation 2001 (IES 2000 empirical distribution)	0.822	0.817	0.827

Source: Census 1996; Census 2001 (authors' own calculations).
Note: All results were weighted using weights supplied by Statistics South Africa.

To date, investigations into the changes in poverty and inequality between 1996 and 2001 (Leibbrandt *et al* 2004) have not only been hampered by missing income data but by the incomparability of the income bands. In the introduction, we cited Cronje and Budlender's (2004) discussion of this problem. The top income band in 1996 was R30,000 or more. This is lower than the real income equivalent of the top three bands in 2001. In order to compare inequality in 1996 and 2001, Leibbrandt *et al* compressed the top end of the 2001 distribution into the real income equivalent of the top band in 1996. Our best case set of multiple imputations for missing values imputes the missing values into income bands and then uses the empirical distribution of personal income based on the 1995 and 2000 Income and Expenditure Surveys (Statistics South Africa 1995, 2000) to derive intra-band points estimates of inequality. In terms of the comparisons over time, this has a major advantage because we do not have to restrict ourselves to comparable income bands as we are able to model the distribution of incomes in the upper bands for 2001 and in the unbounded highest band for both periods. Our estimates of changes in poverty and inequality in Tables 6 and 7 not only take into account the potential bias an uncertainty due to missing values but also use the full distribution of incomes.

Table 6 and 7 present the resultant comparisons of poverty and inequality in 1996 and 2001. Table 6 presents estimates of the poverty headcount for both a lower and higher poverty line. While there has been a small increase in poverty measured at the lower line, the increase at the higher line is more marked with almost 8% more individuals finding themselves in poverty in 2001 than in 1996. Turning to the poverty gap ratios in Table A3 of the Appendix, we also see a marked increase in the poverty gap ratio indicating that the depth of poverty has also increased over the period. Table 7 presents a comparison of Gini coefficients in 1996 and 2001. Our results suggest a marked increase in inequality over the period.

Thus, at the end of a lot of careful imputation work on the ten percent micro sample of the 1996 and 2001 census, our results confirm the major findings from the existing literature (Leibbrandt *et al* (2004)) in that we find small increases in poverty for the poorest of the poor between 1996 and 2001, more marked increases when a higher poverty line is used and unambiguous increases in inequality. For both 1996 and 2001, our estimates of the poverty and inequality parameters are combined estimates embodying five sequential regression imputations for missing values. As such, we would argue that they are the currently best available estimates of these parameters.

References

Babita, M., Demombynes, G., Makhatha, N., Ozler, B., (2002). “Estimated Poverty and Inequality Measures in South Africa: A Disaggregated Map for 1996”. Unpublished manuscript.

Cronje, M. and Budlender, D. (2004) “Comparing Census 1996 and Census 2001: An operational perspective.” *South African Journal of Demography* 9(1) 67-89.

Foster, J.E., Greer, J and Thorbecke, E (1984) A class of decomposable poverty indices, *Econometrica*, 52, 761–766.

Leibbrandt, M., Poswell, L., Naidoo, P., Welch, M. and Woolard, I. (2004) “Measuring Recent Changes in South African Inequality and Poverty Using 1996 and 2001 Census Data” *CSSR Working Paper* 84, Centre for Social Science Research, University of Cape Town.

Little, R.J. and Rubin, D.B. (2000) *Statistical Analysis with Missing Data*, Wiley, New York.

Lohr, S.L. (1999) *Sampling: Design and Analysis*, Pacific Grove: Brooks/Cole Publishing Company.

McGrath, M.D. (1983) “Inequality in the Size Distribution of Personal Income in South Africa in Selected Years Over the Period from 1945 to 1980”, Unpublished PhD. Dissertation, University of Natal Durban.

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J. and Solenberger, P. (2001) “A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models.” *Survey Methodology* 27(1) 85-95.

Simkins, C. (2005) “What Happened To The Distribution Of Income In South Africa Between 1995 And 2001?”. Unpublished paper. University of the Witwatersrand.

Statistics South Africa (1995) *Income and Expenditure Survey 1995*, Pretoria: Statistics South Africa.

Statistics South Africa (1996) *Census 1996 10% Sample*, Pretoria: Statistics South Africa.

Statistics South Africa (1997) Consumer Price Index (CPI), *Statistical Release P0141.4* Statistics South Africa, Pretoria.

Statistics South Africa (2000) *Income and Expenditure Survey 2000*, Pretoria: Statistics South Africa.

Statistics South Africa (2001a) *Census 2001 10% Sample*, Pretoria: Statistics South Africa.

Statistics South Africa (2001b) “Consumer Price Index (CPI)” *Statistical Release P0141.1. October 2001* Pretoria: Statistics South Africa.

Van Buuren, S., Boshuizen, H.C. and Knook, D.L. (1999) “Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis” *Statistics in Medicine* 18 pp. 981-694.

Whiteford, A.C. and McGrath, M.D. (1994) *Distribution of Income in South Africa*, Pretoria: Human Sciences Research Council.

Whiteford, A.C. and van Seventer, D.E. (2000) "South Africa's changing income distribution in the 1990s" *Journal of Studies in Economics and Econometrics* 24(3) 7-30.

Appendix

Table A1 Rates of missing data in variables used in SRMI Multiple Imputation (un-weighted numbers and percentages)

	Number of individuals with missing data	*Percentage of sample with missing data*
Age	25,976	0.72%
Gender	45,358	1.26%
Employment	196,918	5.47%
Occupation	218,855	6.08%
Education	236,578	6.57%
Income	561,095	15.59%

Source: Census 2001 (authors' own calculations).

Table A2 Rates of missing income data by other variables used in SRMI Multiple imputations (un-weighted numbers and percentages)

	Number of individuals in category with missing income data	*Percentage of category with missing income data*
Race		
Black African	400,454	14.01
Coloured	69,252	20.85
Indian/Asian	15,822	17.19
White	75,567	23.8
Location		
Rural	205,065	13.13
Urban	356,030	17.47
Province		
Western Cape	86,006	23.38
Eastern Cape	99,473	18.39
Northern Cape	8,675	12.76
Free State	32,968	15.11
KwaZulu-Natal	111,707	15.61
North West	20,761	6.94
Gauteng	131,930	19.01
Mpumalanga	27,492	10.77
Limpopo	42,083	9.53
Age		
< 11	189,848	22.27
11- 20	150,689	18.76
21 - 30	82,495	13.61
31 - 40	50,134	10.32
41 – 50	35,849	10.06
51 – 60	23,608	10.76
> 60	21,397	8.55

	Number of individuals in category with missing income data	*Percentage of category with missing income data*
Gender		
Male	252,492	15.13
Female	296,450	15.72
Employment Status		
Employed	36,959	5.26
Not employed	438,229	16.23
Occupation		
Legislators, senior officials and managers and professionals	6,125	6.61
Elementary occupations	5,471	2.91
Other occupations	19,662	4.92
Education		
No education	38,386	9.4
Less than Grade 7	136,131	15.37
Less than Grade 12	149,425	13.81
Grade 12	57,242	13.59
More than Grade 12	16,376	10.3

Source: Census 2001 (authors' own calculations).

Note: Some of the variables have missing values (see Table A1) so the denominator is not consistent throughout the table.

Table A3 Poverty Gap Ratios 1996 and 2001

	Poverty Gap Ratio (Poverty line at R124 per capita per month)			*Poverty Gap Ratio (Poverty line at R400 per capita per month)*		
	Estimate	*95% C.I.*		*Estimate*	*95% C.I.*	
No imputed values 2001	0.321	0.320	0.321	0.514	0.514	0.515
Statistics South Africa's hot deck imputation 2001	0.295	0.294	0.295	0.485	0.484	0.485
SRMI Multiple imputation 2001 (mid-points)						
Combined	*0.297*	*0.286*	*0.309*	*0.487*	*0.476*	*0.498*
Imputation 1	0.289	0.289	0.290	0.479	0.479	0.480
Imputation 2	0.299	0.298	0.299	0.488	0.488	0.489
Imputation 3	0.304	0.303	0.304	0.493	0.493	0.494
Imputation 4	0.298	0.298	0.299	0.488	0.488	0.489
Imputation 5	0.298	0.297	0.298	0.488	0.487	0.488
SRMI Multiple imputation 2001 (implausible values set to missing)						
Combined	*0.223*	*0.125*	*0.321*	*0.444*	*0.366*	*0.522*
Imputation 1	0.219	0.219	0.220	0.444	0.444	0.444
Imputation 2	0.260	0.260	0.261	0.471	0.471	0.472
Imputation 3	0.233	0.233	0.233	0.454	0.454	0.455
Imputation 4	0.147	0.147	0.148	0.382	0.382	0.382
Imputation 5	0.256	0.255	0.256	0.468	0.467	0.468
SRMI Multiple imputation 2001 (IES 2000 empirical distribution)						
Combined	*0.302*	*0.292*	*0.311*	*0.492*	*0.480*	*0.503*
Imputation 1	0.299	0.299	0.300	0.491	0.491	0.492
Imputation 2	0.299	0.298	0.299	0.488	0.487	0.488
Imputation 3	0.304	0.303	0.304	0.493	0.492	0.493
Imputation 4	0.308	0.307	0.308	0.500	0.500	0.501
Imputation 5	0.298	0.298	0.299	0.487	0.487	0.488
SRMI Multiple imputation 1996 (IES 1995 empirical distribution)						
Combined	*0.272*	*0.271*	*0.273*	*0.421*	*0.420*	*0.422*
Imputation 1	0.272	0.271	0.272	0.421	0.420	0.421
Imputation 2	0.273	0.272	0.273	0.422	0.421	0.422
Imputation 3	0.271	0.271	0.272	0.420	0.420	0.421
Imputation 4	0.272	0.271	0.272	0.421	0.420	0.421
Imputation 5	0.272	0.272	0.273	0.421	0.421	0.422

Source: Census 1996; Census 2001 (authors' own calculations).
Note: All results were weighted using weights supplied by Statistics South Africa.

The Centre for Social Science Research

Working Paper Series

RECENT TITLES

Behar, A. 2004. *Estimates of labour demand elasticities and elasticities of substitution using firm-level manufacturing data*. Cape Town. CSSR Working Paper No. 98.

Simchowitz, B. 2004. *Social Security and HIV/AIDS: Assessing "Disability" in the Context of ARV Treatment.* Cape Town. CSSR Working Paper No. 99.

Mills, E. 2004. *Beyond the Disease of Discrimination: A Critical Analysis of HIV-Related Stigma in KTC.* Cape Town. CSSR Working Paper No. 100.

du Toit, A. 2005. *Forgotten by the highway: Globalisation, adverse incorporation and chronic poverty in a commercial farming district.* Cape Town. CSSR Working Paper No. 101.

Perkins, P. Fedderke, J. Luiz, J. 2005. *An Analysis of Economic Infrastructure Investment in South Africa.* Cape Town. CSSR Working Paper No. 102.

Fedderke, J. Perkins, P. Luiz, J. 2005. *Infrastructural Investment in Long-run Economic Growth: South Africa 1875-2001.* Cape Town. CSSR Working Paper No. 103.

Seekings. J. 2005. *Prospects for Basic Income in Developing Countries: A Comparative Analysis of Welfare Regimes in the South.* Cape Town. CSSR Working Paper No. 104.

Fedderke, J. Szalontai, G. 2005. *Industry Concentration in South African Manufacturing Industry: Trends and Consequences, 1972-96.* Cape Town. CSSR Working Paper No. 105.

The Centre for Social Science Research

The CSSR is an umbrella organisation comprising five units:

The Aids and Society Research Unit (ASRU) supports quantitative and qualitative research into the social and economic impact of the HIV pandemic in Southern Africa. Focus areas include: the economics of reducing mother to child transmission of HIV, the impact of HIV on firms and households; and psychological aspects of HIV infection and prevention. ASRU operates an outreach programme in Khayelitsha (the Memory Box Project) which provides training and counselling for HIV positive people

The Data First Resource Unit ('Data First') provides training and resources for research. Its main functions are: 1) to provide access to digital data resources and specialised published material; 2) to facilitate the collection, exchange and use of data sets on a collaborative basis; 3) to provide basic and advanced training in data analysis; 4) the ongoing development of a web site to disseminate data and research output.

The Democracy in Africa Research Unit (DARU) supports students and scholars who conduct systematic research in the following three areas: 1) public opinion and political culture in Africa and its role in democratisation and consolidation; 2) elections and voting in Africa; and 3) the impact of the HIV/AIDS pandemic on democratisation in Southern Africa. DARU has developed close working relationships with projects such as the Afrobarometer (a cross national survey of public opinion in fifteen African countries), the Comparative National Elections Project, and the Health Economics and AIDS Research Unit at the University of Natal.

The Social Surveys Unit (SSU) promotes critical analysis of the methodology, ethics and results of South African social science research. One core activity is the Cape Area Panel Study of young adults in Cape Town. This study follows 4800 young people as they move from school into the labour market and adulthood. The SSU is also planning a survey for 2004 on aspects of social capital, crime, and attitudes toward inequality.

The Southern Africa Labour and Development Research Unit (SALDRU) was established in 1975 as part of the School of Economics and joined the CSSR in 2002. SALDRU conducted the first national household survey in 1993 (the Project for Statistics on Living Standards and Development). More recently, SALDRU ran the Langeberg Integrated Family survey (1999) and the Khayelitsha/Mitchell's Plain Survey (2000). Current projects include research on public works programmes, poverty and inequality.