RESEARCH NOTE: ERRORS IN THE OCTOBER HOUSEHOLD SURVEY 1994 AVAILABLE FROM THE SOUTH AFRICAN DATA ARCHIVE[1]

MARTIN WITTENBERG*

Abstract
The on-line electronic documentation supplied with the 1994 October Household Survey by the South African Data Archive (SADA) appears to be incorrect. In particular, the electronic version of the questionnaire does not correspond to the hard copy in the possession of the author. The most serious error is that the race classification in the electronic copy is different from the classification on the hard copy. Researchers relying on the electronic copy will erroneously interchange the categories "Coloured", "White" and "Black". This could lead to seriously misleading analyses. The reason for this mistake can probably be attributed to a retyping of the questionnaire using the 1993 OHS as a template.
JEL Classification: C800
Keywords: October Household Survey 1994; October Household Survey 1993; October Household Survey 1995; South African Data Archive; data quality

The easy availability of electronic data since the advent of democracy has immensely improved the quality and quantity of microeconomic analyses using South African data. The greater availability of data, however, does not always mean better quality control on the information churned out. Indeed as the volume has increased it is quite possible that some data sets in the public domain are subject to much less scrutiny than others. The purpose of this note is to alert the research community to a potentially fatal error in the 1994 October Household Survey accessible through the South African Data Archive.

Statistics South Africa is the biggest producer of data sets, but since the year 2000 the South African Data Archive (SADA) has become, perhaps, the easiest source of these data sets. There are several reasons for this. Firstly, the interface offered by SADA is significantly more user friendly. It is possible to order data sets through a web based form. The data itself will be transferred either by FTP or mailed on a CD. The cost to academic researchers is essentially zero. By contrast, the Statistics South Africa web-page does not indicate how one would go about ordering the data. One needs to contact one of the user consultants in order to gain access.

Secondly, Statistics South Africa data sets arrive in ASCII format. In order to use the data one has to convert these files into a format that one's favourite statistical package will be able to work with. Since the records are arranged in fixed length format one has to manually specify which columns should be read into which variable. This involves significant work. The SADA files, by contrast, arrive in the format of one's choice.

* School of Economics and SALDRU, University of Cape Town.
[1] I would like to thank Lynn Woolfrey of Data First for access to the data and research assistance. I also received assistance from Martine Mariotti at UCLA. This work has been partially supported by a grant from the Mellon Foundation.

Table 1. Codes in the electronic and hard copy of the 1994 OHS questionnaire

Codes	Electronic copy of the questionnaire	Hard copy of the questionnaire	Sample proportions (person file)	Population estimates (weighted sample)
1	Asian	Asian	10,404	1,038,851
2	Black	Coloured	24,412	3,472,178
3	Coloured	White	20,580	5,192,498
4	White	Black	77,073	30,613,467

Table 2. Race codes in the 1993-1995 OHSs

Codes	1993 OHS	1994 OHS	1995 OHS
1	Asian	Asian	African/Black
2	Black	Coloured	Coloured
3	Coloured	White	Indian/Asian
4	White	Black	White

Thirdly, the main search engines hit on SADA and not on the Statistics South Africa web site if the terms "October Household Survey" or "October Household Survey 1994" are entered. Interestingly enough, the first site offering access to the data is the UCLA Institute for Social Science Research Data Archives. However they provide only a mirror of the SADA data sets.

Given this apparent centrality of SADA to the dissemination of South African data sets, it is somewhat disconcerting that there are a number of problems with their version of the 1994 October Household Survey. The most fundamental problem is that the electronic version of the questionnaire does not correspond perfectly to the hard copy in the possession of the author.

The most critical divergence comes in Question 2.1, which asks for the population group of each member of the household. As can be seen from Table 1 the electronic copy is significantly different from the hard copy. Furthermore the sample proportions and the total population estimates clearly indicate that the hard copy is correct and the electronic copy wrong. Indeed the Stata version of the SADA translation of the 1994 OHS "person" file has the codes labelled in line with the hard copy and not with their own electronic codes.

If all the data had been labelled, this mistake would not be all that serious. Regrettably, however, the "worker" file is *not* labelled. Consequently anyone working on that file and referring back to the electronic version of the questionnaire would be seriously led astray. Indeed this author was puzzling over some serious anomalies in the African employment rates in 1994 and became convinced that the race codes had to be wrong.

How serious a problem is this likely to be in reality, given that most analysts have ignored the 1994 OHS and based their discussion of post-apartheid trends beginning with the 1995 OHS? The UCLA web site mentioned above had 50 hits on the 1994 OHS. This suggests that there has been at least some interest by foreign academics in this data set. Perhaps if the labour market information did not look so weird there would have been more!

How could such a mistake have possibly arisen? The easiest explanation is that SADA retyped a hard copy of the 1994 questionnaire to accompany their electronic posting. Indeed the MS Word version of the questionnaire available on the web site was created in the year 2000 at SADA. It is likely that the 1993 OHS questionnaire was used as template. The race codes in that questionnaire were different from those used in 1994, or

indeed, from those used in 1995 as shown in Table 2. This is one of many traps for unwary researchers!

There are additional problems with the SADA data set. The age and education variables in the "worker" file have been truncated to one significant digit. Consequently education has become a binary variable (zero or one) while age has a maximum value of 9. None of these problems exist in the raw data that the author obtained from Statistics South Africa.

The fundamental point is that organisations that reprocess raw data often provide a useful service; but some times can corrupt the underlying information.