SURVEY OF SOCIOECONOMIC OPPORTUNITY AND ACHIEVEMENT IN SOUTH AFRICA

CODEBOOK

Principal Investigators:

Donald J. Treiman, UCLA
Sylvia N. Moano
Lawrence Schlemmer

Original codebook and data prepared by:

Donald J. Treiman, UCLA
Alisa Lewin, Haifa University
Yao Lu, Columbia University
Matthew McKeever, Haverford College
(Lewin, Lu, and McKeever were formerly graduate students in Sociology at UCLA)

Updated codebook and data prepared at Data First, University of Cape Town, South Africa in 2016-2017 by:

Donald J. Treiman, UCLA
Gabriel Espi-Sanchis, DataFirst
Martin Wittenberg, UCT, DataFirst
Andrew Kerr, DataFirst

Note: **It is imperative that any errors or omissions in the data or documentation discovered by users be reported to Lynn Woolfrey, Data First, University of Cape Town, via email [lynn.woolfrey@uct.ac.za] in complete detail, as soon as they are discovered.**

TABLE OF CONTENTS

I. Documentation

II. Variables in Questionnaire

III. Constructed Variables

1. Occupation variables
2. Demographic variables

3. Education variables
4. Income variables

Appendices

Appendix A: Response Categories Used for Many Variables

Codes for language
Codes for highest level of education attained
Codes for main activity
Codes for type of place
Codes for country of birth
Codes for month

Appendix B: Occupation Data

Introduction
Enhanced *International Standard Classification of Occupations* (Revised Edition 1988)
Modifications of standard ISCO-88
Rationale for including specific additional categories
Codebook for "sa_occ_codes_90604.txt"

Appendix C: Geographic Place Codes

Codes for Magisterial Districts
Codes for Enumerator Areas
Codes for Planning Regions
Codes for Cities/Towns
Codes for Suburbs/Townships

Appendix D: Questionnaires

English/South Sotho/Zulu
Afrikaans

Appendix E: -do- and -inc- Files Used to Create the Variables in Part 3

Appendix F: Other Documents

Sampling Memorandum of agreement between MarkData and the principal investigators
Sample specifications and weights
Sampling specifications
Instructions
Fieldwork reports
Social Stratification Report (missing)
Fieldwork Report - Project: Social Stratification (partial copy available)

Appendix G: Bibliography of Publications Cited in this Codebook

I. Documentation

Introduction. This survey was conducted by MarkData, the in-house survey organization of the Human Sciences Research Council (HSRC), Pretoria, Republic of South Africa, for a research team headed by Donald J. Treiman, UCLA; Sylvia N. Moeno, then in the Strategic Planning Group at Eskom, Johannesburg; and Lawrence Schlemmer, then Director of the Centre for Policy Studies, School of Business, University of the Witwatersrand, Johannesburg. The survey was funded by the U.S. National Science Foundation (SES89 12677), and by several South African agencies: the Anglo-American/De Beers Chairman's Fund, the Trust for Educational Advancement in South Africa, the Human Sciences Research Council, Johannesburg Consolidated Investments, and the Union Carbide Corporation.

Field work. Because at the time of the survey, South Africa had for many years practiced a policy of residential segregation by race (authorized under the the "Group Areas Act", which was abolished in June, 1991), South African surveys typically were conducted by drawing separate samples from each of the "Group Areas," that is, areas designated for residence by Whites, Asians, Coloureds, and Blacks.[1] In South Africa, interviewing usually is carried out by interviewers of the same race as interviewees. To accommodate scheduling considerations, the main survey was carried out in two parts: 3,679 interviews were conducted in *W*hite, *A*sian, and *C*oloured areas in April and May 1991 (the WAC sample) and 3,689 interviews were conducted in Black areas in August and September 1991 (the Black sample). Processing of the two parts was also carried out separately. Because there proved to be substantial under-representation of males in both samples (see below), data were collected for a supplementary sample of 749 men in May-July 1992. Finally, data were collected from a special sample of 969 rural respondents in 1993-94 (see below). In all, between 1991-94 data were collected for 9,086 respondents.

Population and samples. The population surveyed consisted of persons age 20 or older residing in "greater South Africa," that is, the Republic of South Africa plus the—at the time—nominally independent "TVBC states": Transkei, Venda, Bophutatswana, and Ceskei.[2] The sample, as designed, was a complex area probability sample, in which sampling categories and the number of cases to be interviewed were determined by the researchers, households were to be selected within areas at random with probability proportionate to size, and adults within households were to be selected at random, but in such a way that the probability that males would fall into the sample was twice the probability that females would fall into the sample. The sample was designed so that, suitably weighted, it could be regarded as representative of the entire adult population of "greater South Africa." The initial sample design was as follows:

<div align="center">

Number of cases targeted

</div>

White areas	2,400
White non-farm areas	2,200
White farms	100

---

[1]   The politics of language in South Africa is complex and constantly in flux. When the survey was conducted, beginning in 1991, the offical racial categories (e.g., those appearing in official publications, including the South African Census) were "White," "Asian," "Coloured," and "Black." Many of the questions in the survey use these terms to make distinctions by race, and specifically use the term "Black" to identify the indigenous Bantu-speaking peoples who make up about three-quarters of the population of South Africa. Although there have been changes in terms of reference to these groups since the survey was conducted, for the sake of clarity we retain the language used in the survey.

[2]   It should be noted that at the time of the survey there also were six "Homelands" or "Self-governing States" (Gazankulu, KaNgwane, KwaNdebele, KwaZulu, Labowa, and Qwa-Qwa) which were regarded as part of the Republic of South Africa, but which were in some instances treated differently from "White South Africa" by the Central Statistical Services, necessitating the creation of special codes by the HSRC.

Black households on White farms		100
Asian non-farm areas	700	
Coloured non-farm areas	700	
Black areas	3,700	
Black areas in "White South Africa"		2,500
TBVC States and homelands [urban and rural]	600	
Black elite areas		400
Black mine hostel dwellers		100
Black township hostel dwellers	100	
Total	7,500	

In practice, there proved to be substantial difficulty in implementing the sample design, and it was modified in the course of the field work.

Male-deficit. First, many interviewers apparently did not correctly implement the modified Kish-table procedure designed to select respondents within households, and generated a sample with too few men and too many women. Fortunately, because information was collected on all residents of a household (Q1.1), it is possible to estimate precisely the sample bias with respect to sex that resulted from the first wave of interviewing.

	Number of males expected	Number interviewed	Difference
White areas	1,409	1,154	255
Asian areas	459	397	62
Coloured areas	435	318	117
Black areas	?	?	?

HSRC went part way toward correcting this bias by conducting at its own expense a supplementary survey of 250 men in the WAC areas and 500 men in the Black areas, in May-July 1992. The actual number of supplementary interviews was 749, as follows:

White areas	150
Asian areas	36
Coloured areas	64
Black areas	499

Departures from strict probability sampling. Second, it appears that in some of the Black urban areas strict probability procedures were not followed, since well over 90 per cent of interviews were completed on the first attempt, a number that is far higher than is typical of other countries and far higher than was the case in the WAC sample, for which less than 75 per cent of interviews were completed in the first attempt. Here are the details (the percentage distribution is shown for the final sample—see below for modifications in the sample design):

Sample (interview date)	Percentage of interviews completed on first attempt
White Areas	
General population (1991-92)	72.5
White Commercial farmers (1991)	100.0
White small holders (1994)	74.2

Sample (interview date)	Percentage of interviews completed on first attempt
Coloured farm laborers on White farms (1994)	98.9
Black farm laborers on White farms (1991)	98.0
Asian Areas	
General population (1991-92)	71.6
Coloured Areas	
General population (1991-92)	71.1
Black areas	
General population - urban (1991-92)	93.6
Rural (homelands and TVBC States) (1994-95)	98.6
Special samples of Blacks	
Black elite areas (1991)	96.2
Township hostel dwellers (1991)	97.9
Mine hostel dwellers (1991)	100.0

Omission of rural sample. Third, due to a misunderstanding between the Principal Investigators and the HSRC sampling staff, the original sample included no persons residing in rural areas, with the exception of a sample of 100 White farmers and 100 Black farm laborers residing on the chosen White farms (however, urban areas in the Homelands and TVBC states were included in the sampling frame for the general population Black sample). To correct this problem, and yield a sample that could be regarded as representative of all of South Africa, a supplementary sample of 1,000 households in rural areas was drawn and field work was conducted in 1993-94.[3] The sample design includes:

	Targeted sample	Realized sample
Rural Blacks in Homelands and TVBC states	800	789
Coloured farm laborers	100	91
White small holders	100	89
Total	1,000	969

Realized sample. The total sample available for analysis thus consists of the following subgroups. These groups are identified in the data set by the variable SAMPLE. Here are the unweighted frequencies, by RACE and SAMPLE. We present this cross-tabulation to indicate the relationship between areas sampled and the racial composition of the sample. As will be discussed below, information on the race of individuals is used when weighting the data.

Sample	Whites	Asians	Coloureds	Blacks	Total
White Areas					
General population	2268	14	10	26	2318
White farmers	98	0	2	0	100
White small holders	88	1	0	0	89
Coloured farm laborers	0	0	91	0	91
Black farm laborers	0	0	0	100	100
Asian Areas					
General population	1	732	13	0	746
Coloured Areas					

---

[3]   In the rural fieldwork, unlike the earlier fieldwork, no attempt was made to choose males at twice the rate of females. The gender composition of both the rural and urban samples are brought into line with the population distribution via post-fieldwork weighting (see below).

Sample	Whites	Asians	Coloureds	Blacks	Total
General population	8	9	745	3	765
Black areas					
General pop. - urban	2	1	7	3512	3522
Rural	0	0	0	789	789
Special samples of Blacks					
Black elite areas	1	1	7	363	372
Township hostel dwellers	0	1	0	96	97
Mine hostel dwellers	0	0	0	97	97
Total	2466	759	875	4986	9086

Under-representation of Black servants living on White "stands." It was common for White householders to employee household domestic servants who lived in separate dwellings in the backyards of the White property or "stand." These dwellings were supposed to be counted as separate households and included in the sample in White areas. But from inspection of line 1 of the table above, it appears that this was not done or was seldom done, since only about 1% of the general population sample in White Areas was Black. Nothing can be done to correct this under-representation of Blacks in White Areas, but it should be noted.

Weighting
written by Gabriel Espi-Sanchis (Data First, University of Cape Town)

This section describes the weights contained in the data set, the subgroups of the sample and population that they apply to, and the variables that can be analysed using them.

As discussed above, the data set is sorted into various subgroups by the variable *sample*, each of which can be considered a separate sample collected through a distinct process. These can be thought of as strata and are labelled as follows: white urban, asian urban, coloured urban, black urban, black rural, commercial farm, smallholder area, coloured farm labourers (on white farms), black farm labourers, black elite area, urban hostels, and mine hostels.

For the purposes of weighting (and distinguishing between samples more faithfully) an expanded sample variable, x*sample,* was created that separated each of the four urban samples into two groups: the original sample collected in 1991 and the additional male sample collected in 1992. As discussed earlier, the Kish tables used for the original urban samples were meant to sample males at a rate twice as high as that for females. However, these seem to have been disregarded, resulting in only 46.89% of respondents being male, and so additional males were sampled in the following year (150 white, 36 asian, 64 coloured, and 498 black respondents). The rationale for sampling more males than females was that males would yield more valuable labour market and occupational information, although the consequent complexities mean this appears to have been a mistake. However, the additional males have similar prior weights to the original samples, are drawn from a wide spread of provinces, magisterial districts and EAs, and are described as being

probability samples in the documentation and correspondence, and so it appears worthwhile to find a way to analyse them in conjunction with the original samples.

We have created four sets of weights: *ceweight2_hh*, *ceweight2_ind, ceweight2m_hh* and *ceweight2m_ind.* The first two are based on the four original urban samples (in 1991) and the black rural sample of 1993/1994 (henceforth collectively referred to as the 'original sample'). These are the samples that we are most certain are probability samples and for which the creation of weights was most straightforward and involved the least assumptions. The latter two are based on the same sample but with the further inclusion of the additional males (henceforth collectively referred to as the 'expanded sample'). The fact that many households and male respondents could theoretically have been drawn in either the original or additional sample meant that the creation of these weights relied on the creation of adjustment factors and some significant assumptions. For both the original sample and the expanded sample, both a household weight and individual weight are included. As yet we have not defined any weights for the remaining (so-called special) samples, principally for three reasons. First, the documentation and correspondence includes a note that states that these cannot be considered probability samples. Second, the weight we consider a trustworthy foundation (see below) is not defined for most of the special samples. Third, for many of the special samples, such as the black elite and hostel dwellers, there is a positive probability that some members of these groups are already included in the populations from which the other samples are drawn, but there is no sure way to determine which or how many respondents (and thus no safe way to down-weight them to attenuate the effect of double counting). On the other hand, those special samples that are definitely not included in any of the originally sampled populations, such as white farmers and coloured farmworkers on their farms, could be incorporated into future weights.

These weights are all calculated using preexisting weights from the data, *wt1*, as the foundation. *wt1* is defined, with a few missing values, for all samples except for white farmers, black farmworkers, all hostel dwellers and the black elite. We assumed that these weights reflected important elements of the sample design, including differential rates of sampling in different magisterial districts and provinces. This assumption was based on the observations that they could be almost perfectly predicted by the magisterial districts (with an R-squared of 0.9968 for Africans in the original sample) and that several estimates, for example household size, looked very strange when they were not used. It is not clear to what extent these weights have been calibrated to population variables, what the source of these benchmarks was, and which variables were used. In order to calibrate these weights on some demographic characteristics (such as age and gender) and enable more direct comparison of the data with other surveys, we used data from the 1991 Census.

Before attempting to isolate the equivalent population in the Census and calculate totals and proportions for calibration, it is necessary to make explicit what the target populations of the SSOA samples were. In the original sample the target population was the urban population of all four races, and the rural black population. Any household in these areas with a member age 20 years or more could have been sampled. However, the additional male sample would have had a slightly different target population (at the household level), given that any household which didn't have any adult males could not reasonably have been sampled (and indeed there are none in the data). Therefore this sample will need to be calibrated to a different population excluding all-female households.

The structure of the SSOA is that there is one adult respondent (of 20 years or more) per household for whom a wealth of information is available, and minimal information (covering age, sex, race and relation to household head) is available for the rest of the household. Therefore, to calibrate the sample to some external population estimates there are two options: to calibrate the respondents to the equivalent population of adults, or to reshape the

data set so that each roster member is its own observation and then calibrate this to the equivalent population of all ages. We took the second option on the reasoning that a calibration based on several times more observations would be more likely to yield reliable and well-fitted household weights.

The 1991 Census was carried out in South Africa excluding the nominally independent homelands (TBVC states: Transkei, Bophuthatswana, Venda and Ciskei) but separate Censuses were performed in these areas. We had access to the actual data for all of these except Transkei, for which we had population estimates and which we assumed were all rural. For the first calibration, we calculated the proportions from the 1991 Census, by gender and in 10-year age brackets, of people in each of the five subgroups of the original samples. An urban/rural variable and race variable allowed us to isolate the proportions for the four races in urban populations and for the black rural population. We created a variable to identify households without any adult males, and these were omitted from the four urban populations for the second calibration (which also excluded the black rural population).

After reshaping the SSOA and creating the same proportions for 10-year age brackets by sample and gender, these were calibrated by Takwanisa Machemedze (Data First) so that the proportions were similar to those of the Census but still reflected the prior weights *wt1*. This gave us a household weight, *ceweight2_hh,* defined for 7,392 observations. To convert this into an individual weight one needs to multiply it by the inverse of the probability of intra-household selection. Given the fact that the Kish tables seem to have been ignored we assume that all adults in the household were sampled with equal probability, giving us a probability of selection of 1/no. adults in hh. Multiplying the household weights by the inverse of this gives us *ceweight2_ind*.

The question of how to incorporate the additional males was far more complex. Once again, we decided that calibrating to a reshaped household data set would constitute a more accurate calibration, meaning that the target population was not just adult males in urban areas but all people in household with at least one adult male. To perform the calibration we had to initially act as if the additional sample was the entire sample. As there were only 36 Asian households, the coefficients for the Indian subgroups were too high in the calibration process, and they were omitted, meaning weights are not defined for additional Asian males and their households. This left us with two household weights, both of which added to totals of distinct but largely overlapping samples. In combining these weights we had to treat them as if they were completely separate samples and had to assume that no respondents were sampled twice (luckily the probability of this is tiny). To generate a single household weight it was necessary to distinguish between households that could theoretically have been sampled through both sampling processes, and those that could only have been sampled in the original sample (households with no adult males). The latter group retained their original weight, while all households from the prior group had to be weighted down. The adjustment factor we used for the prior groups was the number of households in the sample (either in the original or the additional sample, respectively) over the sum of households sampled in both the original and additional samples. This yielded the weights *ceweight2m_hh,* defined for 8,104 respondents.

For individuals the process was slightly more complex and a different adjustment factor was needed. The first step in turning the calibrated weights (*not* the household weights, which have been multiplied by the household adjustment factor) into individual weights is multiplying by the inverse of the probability of intra-household selection, as before. Here, however, only males were selected and so the probability of selection is 1/no. of adult males. Now we needed to calculate an adjustment factor to apply to the individual weights (those just calculated and *ceweight2_ind*). Again, those who could only have been selected in the original sample (all women and all black rural respondents) retain their original weight. The

adjustment factor applied to all those who could have been included in both samples was the number of male respondents (either in the original or the additional sample, respectively) over the sum of male respondents sampled in both the original and additional samples. This process yielded the weights *ceweight2m_ind.*

Despite the different processes by which they were created, the two household weights *ceweight2_hh* and *ceweight2m_hh* can both be used for the same variables, as can the two individual weights *ceweight2_ind* and *ceweight2m_ind.* The household level weights can be used with any variable that is constant at the household level, such as current provision of services (including water and toilet access), as well as for the variables which are defined for all members of the household roster (age, race, gender and relationship to household head, but only after reshaping the data file into a long format). The individual level weights can be used with any of the individual level variables, including a host of occupational, employment, demographic and historical variables.

At each step we performed various checks and compared estimates to others from the SSOA weighted by *wt1*, the Census and the 1993 Project for Statistics on Living Standards and Development (PSLSD. The household weights were found to give estimates of population size, access to services and gender and race distributions that appeared plausible alongside estimates from the PSLSD and the 1991 Census. Below is a table comparing individual-level estimates using wt1 (for the expanded sample) and both sets of individual weights to estimates from the PSLSD and Census. It is clear that the created individual weights produce estimates that are in a plausible range, and that in some cases, as with enrolment rates, proportion male, age, years of education labour force participation, and unemployment the demographic calibration has led to estimates that are closer to those of the PSLSD and Census. Given the relatively small number of additional males sampled and the resultantly low adjustment factors these observations were multiplied by, the differences between the estimates of the two individual weights are generally small, although the additional information provided by these males can be seen in the reduced standard errors and has made some difference to the estimates[4]. Ultimately whether to use *ceweight2m* or *ceweight2* depends on whether the user considers the additional information and sample size provided by the additional males to offset the added assumptions and more complex process needed to create the expanded weights.

Table comparing estimates of individual-level variables with varying weights and data sources (standard errors in parentheses).

	wt1	ceweight2_ind	ceweight2m_ind	PSLSD	Census 1991
Sample	8,036	7,385	8,097	37,096	
Estimated population	14,077,671	17,501,905 (1,197,797)	17,313,326 (1,067,288)	18,658,217 (456,767)	17,621,830
Age	39.36	39.11 (.33)	39 (.31)	39.07 (.25)	39.1
Proportion male	.5063	.5053 (.0114)	.4990 (.0102)	.4548 (.0043)	.49

---

[4] It must be noted that the slightly reduced proportion of males in the third column is an artefact of the handful of females who somehow were sampled in the additional male sample.

Table comparing estimates of individual-level variables with varying weights and data sources (standard errors in parentheses).

	wt1	ceweight2_ind	ceweight2m_ind	PSLSD	Census 1991
% Married*	.4999	.5157 (.0110)	.5063 (.0107)	.5382 (.0106)	.4990
Household size*	4.65	5.53 (.11)	5.54 (.1045)	5.67 (.13)	4.52
Years of education*	8.73	8.3 (.13)	8.33 (.12)	7.39 (.13)	7.11
Has matric	.2905	.2688 (.0140)	.2716 (.0129)	.2119 (.0127)	
Enrolment rate*	.0168	.0269 (.0041)	.0259 (.0040)	.0450 (.0026)	
LFP rate	.6764	.6492 (.0109)	.6519 (.0103)	.6458 (.0092)	
Unemployment rate (broad)	.2145	.2462 (.0147)	.2533 (.0137)	.2962 (.0123)	.183 (narrow)
Mean income*	1587.56	1454.21 (61.82)	1462.15 (54.78)	1911.73 (119.31)	
The first three columns use SSOA data, with the weights applied giving the columns their name.					

Language of interview. For the WAC sample, interviewing was conducted in English or Afrikaans, depending on the preference of the respondent. Questionnaires were printed in both languages, and the responses were recorded in the language of the questionnaire. For the Black samples, partly tri-lingual questionnaires were printed. Questions were shown in English, South Sotho (in bold), and Zulu (beginning with *), but response categories were shown only in English. Interviewers were required to have at least a Standard 10 (complete secondary) education. As noted above, interviewers generally were of the same race as respondents. The interviewers for the WAC sample were nominally literate in both English and Afrikaans, and the interviewers for the Black sample were nominally literate in both English and the dominant African language in the area in which the interview took place.

Data preparation. The preparation of data for analysis was an unusually lengthy and complex task. Many errors and anomalies were discovered in the data, which necessitated a lengthy exchange of correspondence between the HSRC staff and the Principal Investigator at UCLA. The HSRC staff made no attempt to edit the open-ended "other (specify)" responses, but simply listed them in the fieldwork reports. These were recoded (to new or existing codes) at UCLA. All of the occupation coding was done at UCLA, and both the occupation and other variables described in Part III were created at UCLA. Finally, the weights described above were created at DataFirst, based on weights created at UCLA.

Other documentation. In addition to this codebook the documentation for this survey included three reports on the field work. The first report, *Social Stratification Report*, provided information on the initial WAC sample but has not been located. This report contains, among other information, a copy of the English language questionnaire used for the WAC sample and responses to open-ended questions in the WAC sample. The second report, *Fieldwork Report - Project: Social Stratification*,

provided information on the Black sample, the special samples, and the supplementary sample of males but has not been located. This report contains, among other information, the weights HSRC applied to the data, a copy of the questionnaire used for the Black sample, and responses to open-ended questions in the Black sample. The third report, *Social Stratification: Rural Sample*, provides information on the rural sample and a revised set of weights for all samples. It also contains the English, Sotho and Zulu questionnaires.

In addition to the file described in this codebook, there is a separate file (named "sa_occ_codes_90604.txt") of the narrative descriptions of jobs. This file, described in detail in Appendix B, contains six variables:

> the ISCO-88 code assigned to the occupation plus a fifth digit for employment status
> the respondent ID number (see the discussion of the ID variable in Part II)
> an alpha code identifying the person to whom the occupation refers: respondent's current job,
>> respondent's next previous occupation, father's occupation, etc.
> verbatim occupational title
> verbatim description of main duties
> verbatim description of the work setting

Organization of codebook and data files. The data file prepared for analysis is a Stata 14 system file. The order of variables in the file generally follows their order in the codebook, and are listed in Part II and III of this codebook. However, variables derived from the rosters are listed column-wise rather than row-wise, in order to facilitate the preparation of Stata -do- files or other sets of analytic commands.

In addition to the questions shown in the questionnaire, a number of constructed variables have been added to the file. These, are shown in Part III. First, we have created variables indicating the age, sex, race, and relation to household head of the respondent, which are constructed from the household roster plus the response to Q. 1.2. Second, all of the education variables have been converted into years of school completed. Third, income from the main job and income from secondary economic activity have been recoded to ratio variables: annual income. Fourth, two numeric variables have been created for each of the 22 occupations included in the questionnaire: codes corresponding to an enhanced version of the *International Standard Classification of Occupations, 1988 Edition* (ISCO-88); and *International Socioeconomic Index of Occupations scores* (ISEI) (Ganzeboom and Treiman 1996). As noted above, the occupation information is described in Appendix B.