



AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
<http://ageconsearch.umn.edu>
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

Technical Paper Series



Technical Paper 2003: 2

**Creating a 1995 OHS and a
Combined OHS-IES Database in
STATA**

*Eisenburg
September 2003*

PROVIDE

PROJECT

The Provincial Decision-making Enabling Project

Overview


The Provincial Decision-Making Enabling (PROVIDE) Project aims to facilitate policy design by supplying policymakers with provincial and national level quantitative policy information. The project entails the development of a series of databases (in the format of Social Accounting Matrices) for use in Computable General Equilibrium models.

The National and Provincial Departments of Agriculture are the stakeholders and funders of the PROVIDE Project. The research team is located at Elsenburg in the Western Cape.


PROVIDE Research Team


Project Leader:	Cecilia Punt
Senior Researchers:	Kalie Pauw Esther Mohube
Junior Researchers:	Benedict Gilimani Lillian Rantho Rosemary Leaver
Technical Expert:	Scott McDonald
Associate Researchers:	Lindsay Chant Christine Valente

PROVIDE Contact Details

 Private Bag X1
Elsenburg, 7607
South Africa

 ceciliap@elsenburg.com

 +27-21-8085191

 +27-21-8085210

For the original project proposal and a more detailed description of the project, please visit www.elsenburg.com/provide

Creating a 1995 OHS and a Combined OHS-IES Database in STATA¹

Abstract

This paper builds on *Technical Paper 2003: 1*. It provides a technical description of how the 1995 October Household Survey (OHS) dataset was created in STATA and subsequently merged with the 1995 Income and Expenditure Survey (IES) dataset to form a combined dataset. The combined survey links household-level income statistics from IES 1995 with individual-level employment data from OHS 1995. This linking of data is necessary for the construction of the Social Accounting Matrix (SAM) sub-matrix that links factor income data to household income. Since the same set of households were visited for both surveys it is possible to merge observations using unique household identification numbers. All the STATA do-files are listed in the technical appendix.

¹ The main author of this paper is Kalie Pauw, Senior Researcher of the PROVIDE Project. For more details on the research team refer to the final page of this report.

Table of contents

1.	OVERVIEW	1
2.	DESCRIPTION OF THE OHS 1995	1
3.	CREATING THE OHS 1995 DATASET AND MERGING IT WITH THE IES 1995 DATASET.....	2
3.1.	<i>Reading in data</i>	2
3.2.	<i>Merging the various datasets</i>	3
3.3.	<i>Do-file households.do and sub-routines</i>	7
4.	SUMMARY.....	11
5.	TECHNICAL APPENDIX	12
5.1.	<i>Variables contained in the OHS 1995 fixed-format files (layout.txt)</i>	12
5.2.	<i>Listing of do-files used to create the combined dataset</i>	17

List of figures

Figure 1:	Structure of the STATA procedure for creating the combined dataset	2
-----------	--	---

List of tables

Table 1:	List of factors.....	4
Table 2:	Merging the OHS 1995 and the IES 1995 datasets	6
Table 3:	Mixed-race households using IES 1995 and OHS 1995 race classification.....	8
Table 4:	Race of head of the household using IES 1995 and OHS 1995 race classification....	8
Table 5:	Mixed-race households using only OHS 1995 race classification	9
Table 6:	Household groups (SAM household accounts).....	10

1. Overview

An important part of the second phase of the PROVIDE Project is to gather detailed information on household demographics, income distribution and poverty. A number of findings are presented in various PROVIDE Project *Background* and *Technical Papers*. The 1995 Income and Expenditure Survey (IES 1995) is useful for analyses on the aggregate income and expenditure side of households from various race groups, settlement areas, and provinces (see *Technical Paper 2003: 1*). In order to gather information on individuals within households the 1995 October Household Survey (OHS 1995) is drawn on. The OHS 1995 contains individual-level data on respondents' perceived quality of life (access to basic facilities, housing, safety and medical care), demographic issues (age distribution, migration, gender, education, health, disabilities, and welfare), as well as employment issues. The OHS 1995 also provides information on the sources of income of individuals, the type of occupations that they are involved in etc. For the purpose of constructing a Social Accounting Matrix (SAM) it is necessary to combine the household level IES 1995 (particularly income and expenditure data) with the individual-level OHS 1995 (particularly data pertaining to the sources of income, e.g. industry employed in and actual wage-income earned).

This paper provides a technical description of how the OHS 1995 dataset was created in STATA and subsequently merged with a shortened version of the IES 1995 dataset to form a dataset named `combined.dta`. The following section gives a brief description of the structure of the OHS 1995 data files. Section 3 explains the function of each of the `do-files` used to control the merge process. All the `do-files` are listed in the appendix.

2. Description of the OHS 1995

Similar to the IES 1995 datasets, the OHS 1995 dataset is stored in a number of large fixed-width format text files (named `house.txt`, `person.txt`, `work.txt`, `death.txt` and `birth.txt`)². The file **house.txt** contains household-level information relating to the first section in the questionnaire. A single person was requested to answer the questions in this section on behalf of the other household members. The data file contains information on the type of dwelling that the household lives in, energy sources used for cooking, lighting and heating, sanitation and access to land. It also contains data about the perceived quality of life, measured in terms of indicators such as crime in the neighbourhood, pollution, medical facilities and general 'feeling' of the household. The file **person.txt** relates to the second section of the

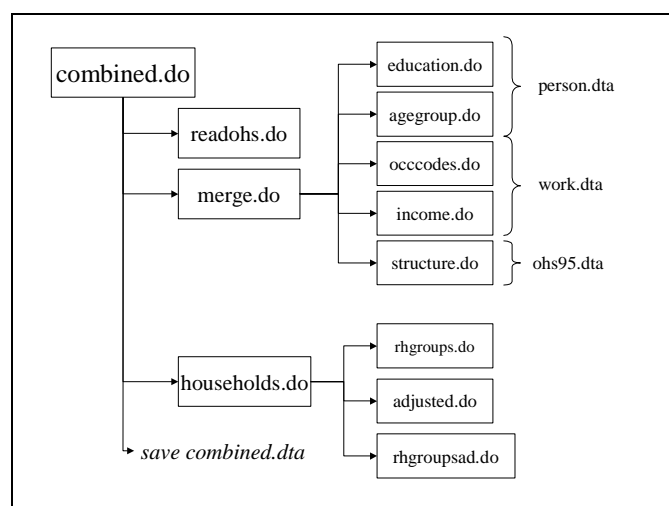
² Note that these files are stored as `*.rsa` files in the original datasets supplied by Statistics South Africa. The files were renamed to `*.txt` files to be consistent with the file extensions used in the creation of the IES 1995 datasets in STATA. Also note that only the first three files were used for the construction of the SAM, as information relating to deaths and births is not required.

questionnaire and was conducted at an individual level. It contains data on demographic measures such as age, gender, race, migration, education, illnesses, and disabilities of each household member. The file **work.txt** provides a detailed overview of each household member's employment status, occupation, the industry in which he or she operates in, as well as individual level income figures.

3. Creating the OHS 1995 dataset and merging it with the IES 1995 dataset

The master do-file **combined.do** runs various sub-do-files, namely **readohs.do**, **merge.do** and **households.do**. **Merge.do** in turn runs five sub-do-files **education.do**, **agegroup.do**, **occcodes.do**, **income.do** and **structure.do**, while **households.do** runs three do-files **rhgroups.do**, **adjusted.do** and **rhgroupsad.do**. The structure of the STATA program that creates **combined.dta** is presented in Figure 1. Each of the sub-do-files is discussed in the following sections of this paper. Do-files are attached in the technical appendix.

Figure 1: Structure of the STATA procedure for creating the combined dataset



3.1. Reading in data

The primary task of do-file **readohs.do** is to read the raw OHS data into STATA. Data from the three OHS 1995 data files **house.txt**, **person.txt** and **work.txt** are read in using the *infile* command. Once all the files have been read in the files are saved as **house.dta**, **person.dta** and **work.dta**. The names of the variables contained in these STATA files relate to the question number, e.g. a variable called *q1_2* in **house.dta** refers to question 1.2 (main type of dwelling). Questions with subsections, e.g. question 1.12, are stored as *q1_12a*, *q1_12b*, etc. The file **layout.txt** is attached in the appendix to this paper and contains a detailed outline of the layout of the data files and the structure of the questions. The OHS 1995 questionnaire is available from Statistics South Africa.

3.2. Merging the various datasets

Do-file **merge.do** controls all the merge processes. Since the combined IES and OHS datasets can become exceptionally large, it is necessary to delete the variables that are not required. Variables are deleted *prior* to merging any two datasets to speed up the merge process in terms of computing time. The user can select variables in each of the datasets that he or she wishes to keep (see various “keep” commands in the file merge.do). Observations are merged using the unique household identification number (*hhid*). Merging is possible since the same set of households were visited for both surveys.

First, house.dta is opened. Variables that won't be needed are deleted and the dataset is saved as houseshort.dta. Secondly, person.dta is opened. Selected variables are deleted and two do-files are run to modify the data in person.dta. Do-file **education.do** groups individuals into groups according to their educational attainment. Do-file **agegroup.do** groups individuals according to the age category (five year intervals) in which they fall. These variables are used in *Background Paper 2003: 3*, which reports on various South African demographics. This paper can be consulted for more information regarding these variables. The dataset is saved as personshort.dta.

Thirdly, work.dta is opened and two do-files are run. The first, **occcodes.do**, reduces the number of categories in the occupation code variable to 11 occupation types. The original dataset has a further breakdown for each of these 11 main occupation groups, but this highly detailed information is not required. The original codes are stored in a variable called *origocc*, while the 11 occupation codes are stored in variable *q3_15*. A variable *occcodes* is also created. This variable is a slight modification of *q3_15*. It combines the occupation code of each individual and his or her race. This variable is made up of a three-digit code where the first digit refers to the race of the individual (1 = African, 2 = Coloured, 3 = Indian and 4 = White), while the second and third digits refer to the occupation code (1 to 11 as per Table 1). An alternative version of this variable is stored in variable *factors*, which simply lists the 44 factor groups (4 race groups multiplied by 11 occupation codes) from 2 through to 45. These are the factor codes used in the South African SAM. Factor code 1 is reserved for the factor capital (see Table 1).

Table 1: List of factors

No.	Factor name
2	African Legislators, senior officials and managers
3	African Professionals
4	African Technicians and associate professionals
5	African Clerks
6	African Service workers and shop market sales workers
7	African Skilled agricultural and fishery workers
8	African Craft and related trades workers
9	African Plant and machinery operators and assemblers
10	African Elementary occupations
11	African Unspecified
12	African Armed forces
13	Coloured Legislators, senior officials and managers
14	Coloured Professionals
15	Coloured Technicians and associate professionals
16	Coloured Clerks
17	Coloured Service workers and shop market sales workers
18	Coloured Skilled agricultural and fishery workers
19	Coloured Craft and related trades workers
20	Coloured Plant and machinery operators and assemblers
21	Coloured Elementary occupations
22	Coloured Unspecified
23	Coloured Armed forces
24	Asian Legislators, senior officials and managers
25	Asian Professionals
26	Asian Technicians and associate professionals
27	Asian Clerks
28	Asian Service workers and shop market sales workers
29	Asian Skilled agricultural and fishery workers
30	Asian Craft and related trades workers
31	Asian Plant and machinery operators and assemblers
32	Asian Elementary occupations
33	Asian Unspecified
34	Asian Armed forces
35	White Legislators, senior officials and managers
36	White Professionals
37	White Technicians and associate professionals
38	White Clerks
39	White Service workers and shop market sales workers
40	White Skilled agricultural and fishery workers
41	White Craft and related trades workers
42	White Plant and machinery operators and assemblers
43	White Elementary occupations
44	White Unspecified
45	White Armed forces

The second sub-do-file, **income.do**, modifies the various income variables. Income is reported as either daily, weekly, monthly or annual. Furthermore, respondents had a choice of

either selecting an income category, ranging from “no income” (code 01) to “R600000 and over” (code 29) or “unspecified” (code 30) (see questions 3.16 and 3.22), or they can report a figure. It is thus necessary to change all income data to annual figures. Persons in each income category were assumed to earn the average category income. Daily income or remuneration was multiplied by 240 (assuming that this is the number of days an individual works per year, weekly income was multiplied by 52 and monthly income was multiplied by 12. Two annual individual income variables are thus created, namely *incmain* (income from main job, working for another person or institution – question 3.16) and *incself* (income from working for oneself – question 3.22). The dataset is saved as *workshort.dta*.

The three data files (*houseshort.dta*, *personshort.dta* and *workshort.dta*) are now ready to be merged using the *merge* command. STATA matches the unique household identification codes (variable *hhid*) contained in each dataset to match observations. This is different from the *append* command used in the creation of the IES 1995 dataset.³ Once the three OHS files have been merged, a do-file **structure.do** is run. This do-file creates a variable that gives a picture of the structure of South African households, grouping each individual within a household into one of 6 categories, namely (1) full-time workers, (2) students, (3), retired or unable to work, (4) unemployed, (5) other and (6) children younger than 15 (note that groups 1 to 5 only include adults over the age of 15). This OHS dataset is saved as *ohs95.dta*.

Finally, the OHS 1995 database is merged with the IES 1995 database (renamed *ies95original.dta* and copied to the OHS 1995 working folder). The resulting file is stored as *combined.dta*. Some problems exist in this combined dataset and need to be dealt with. These issues are discussed below:

- 1) When two files are merged STATA one effectively merges a master dataset with a secondary dataset by linking observations with similar identification codes (*hhid*). During the procedure STATA creates a reporting variable called *_merge*. This variable takes on a value of one for observations in the merged file that were only found in the master data, i.e. the observation originally came from the master data but no matching observation with the same *hhid* could be found in the secondary dataset. A value of two indicates those observations in the merged file that were only found in the secondary dataset. If *_merge* equals three the observations were from both the master and using data. A perfect one-to-one match will have a *_merge*

³ The *append* command adds new observations (sorted by their household identification code) to a dataset, while *merge* adds new variables to existing observations in a dataset. Since the IES 1995 dataset was sorted by province, there were no overlapping observations in the set as each new dataset (province) contained entirely new observations. However, the OHS 1995 data files are grouped by the question number, i.e. each data file contains a few variables relating to every single observation (household).

value of three for all variables. Table 2 shows that this was not the case (note *_merge* was renamed *merge3*):

Table 2: Merging the OHS 1995 and the IES 1995 datasets

Merge3	Description	Percent
1	Observations only in master data	0.77
2	Observations only in secondary data	3.42
3	Observations in both datasets	95.82
Total		100.00

Table 2 shows that the merge was a fairly good match, with almost 96% of the data points (observations) in coming from both datasets. Slightly less than 1% of the observations were only contained in the IES 1995 dataset (master dataset in this specific merger), while over 3% of the observations were only in the OHS 1995 dataset. Although the two datasets come from the same sample of the population, the two surveys were not conducted at the same time. Fieldworkers returned to the same households at a later date, but problems were encountered, especially in rural areas or informal settlements (see Orkin and Hirschowitz, 1996) and as a consequence it is not possible to obtain a perfect match. All mismatched observations were dropped to ensure a perfect match. Thus, a total of 5514 observations out of 131797 (4.2%) were dropped.⁴ These individual observations were members of 1115 households.

- 2) The IES 1995 dataset also contained some problem variables relating to summation and reporting errors (see the discussion in *Technical Paper 2003: 1*). These variables were also dropped. A total of 375 (0.3%) observations from 83 households were dropped.⁵ After these observations have been dropped the combined dataset contains 125908 observations and 28502 households.
- 3) A further problem encountered related to the racial classification of households. For some unknown reason households were not necessarily allocated to the same racial group in the two surveys. The racial classification of the IES 1995 survey is done on the basis of the race of the head of the household, while the OHS 1995 racial classification is at an individual level. Although mixed-race households account for some of the discrepancies, there are quite a number of heads of households that

⁴ This option can be exercised by including the command *keep if merge3 == 3* in *merge.do*.

⁵ These “problem” variables are highlighted in the variable called *_problem*. The option of dropping the problem variables can be exercised by including the command *drop if _problem == 1 | _problem == 2 | _problem == 3*, i.e. drop the variables of problem type 1, 2 or 3 (see *Technical Paper 2003: 1*). Note that the number of problem variables are less than those quoted in *Technical Paper 2003: 1* since some of the problem variables have already been dropped in the previous *drop* statement (see point 1).

reported different racial classifications in the two surveys. Dropping these variables was not considered an option. This issue is discussed further in section 3.3.1.

- 4) Weighting is a final matter of concern. When doing a descriptive overview of data it is often useful to expand the figures to a national level. A perfectly representative sample will not require any weighting. However, the sampling procedure is often biased and hence not representative. Information from, for example a population census, should then be used to re-weight the data. There are some concerns regarding the weights used in the Statistics South Africa IES 1995 and OHS 1995 datasets. These weights were based on the 1991 Population Census, but the 1996 Census revealed a very different picture of the South African society than the weights suggested. The PROVIDE Project mainly uses 'unweighted' data as the aim is merely to find income and expenditure coefficients (i.e. patterns of expenditure) rather than actual levels of expenditure.

3.3. Do-file households.do and sub-routines

The formation of representative household groups is important. The household groups formed here are used directly in the household-factors sub-matrix of the South African SAM that records the distribution of factor payments to households. As was the case with factors, households are also classified according to their racial classification. In order to do this a suitable and consistent race variable should be used. The following section discusses some of the concerns raised in point (2) above regarding the population group variables. Section 3.3.2 continues the discussion about the formation of representative household groups.

3.3.1. *Notes regarding the population group variables*

The IES 1995 dataset contains a population group variable called *race*, which contains four elements, namely African, Coloured, Indian and White. Since the particular survey was conducted at a household level, this variable can be interpreted as the race of the head of the household, and not necessarily as the race of all household members. Thus, information on mixed-race households is missing from this survey. Nevertheless, when household groups are formed for use in the SAM the race of the head of the household is used, thus it is perfectly all right to assume that the *race* variable can be used for this purpose.

When the household-level IES 1995 survey is merged with the individual-level OHS 1995 survey the *race* variable is copied over to each individual in the household, irrespective of that individual's race. Thus, even if an individual in the household's race is different from the head of his/her household, the *race* variable will now report the race of the head of the household for that individual. In order to find the correct race of each individual the variable

s3race should be used. This variable originates from the OHS 1995 survey. A tabulation of *race* against *s3race* should therefore point out how many household members are classified under a different race than the head of their household. Based on the table below there appears to be quite a number of mixed-race households in South Africa (see the off-diagonal elements in Table 3).

Table 3: Mixed-race households using IES 1995 and OHS 1995 race classification

		OHS Classification (<i>s3race</i>)				
		African	Coloured	Indian	White	Total
IES Classification (<i>race</i>)	African	89211	688	117	351	90367
	Coloured	359	16179	102	63	16703
	Indian	26	58	3997	42	4123
	White	258	189	44	14224	14715
	Total	89854	17114	4260	14680	125908

The number of off-diagonal elements seems high. Further investigation revealed that there seems to be an inconsistency between the variable *race* and *s3race* when comparing the values for the head of the household (i.e. *s2persno* = 1). From Table 4 it should be clear that some heads of households reported a different population group when the OHS 1995 survey was conducted, thus explaining why there appears to be so many mixed-race households.

Table 4: Race of head of the household using IES 1995 and OHS 1995 race classification

		Head of Household – OHS Classification (<i>s3race</i>)				
		African	Coloured	Indian	White	Total
Head of Household - IES Classification (<i>race</i>)	African	18429	100	17	123	18669
	Coloured	52	3606	18	15	3691
	Indian	5	9	968	12	994
	White	73	40	9	5026	5148
	Total	18559	3755	1012	5176	28502

In order to minimise discrepancies only the OHS 1995 classification will be used to determine the household group racial classification, using the race of the head of the household as reported in *s3race* (for *s2persno* = 1). This OHS 1995 variable was also used for the racial classification of labour (variable *occcodes* or *factors*). Since the IES 1995 racial classification (variable *race*) will not be used any longer a new variable *headrace* is constructed to replace variable *race*. This variable shows the race of the head of the household as reported in the OHS 1995 racial group variable (*s3race*). Since this variable has to be a household-level variable it is the same for each household member, irrespective of whether

that member's race is different from that of the head of the household. This is necessary since the head of the household's race will be used to construct representative household groups, and therefore all members in the household will be classified according to the race of their head. The following commands are used in STATA to create *headrace*:⁶

```
sort hhid s2persno
by hhid: generate headrace = s3race[1]
```

Table 5 tabulates *headrace* against *s3race*. Off-diagonal entries now represent the 'true' number of individuals from mixed-race households. Clearly the numbers of individuals living in mixed-race households is now rather insignificant. An estimated 1.5% of households can be classified as mixed-race households (own calculations from data).

Table 5: Mixed-race households using only OHS 1995 race classification

	OHS Classification (<i>s3race</i>)					
		African	Coloured	Indian	White	Total
OHS Classification (<i>headrace</i>)	African	89563	400	27	21	90011
	Coloured	254	16630	45	25	16954
	Indian	9	42	4172	4	4227
	White	28	42	16	14630	14716
	Total	89854	17114	4260	14680	125908

3.3.2. Constructing household groups

Now that a suitable population group variable has been constructed the household groups can be created. The classification used here follows that of the SAM for the Western Cape (see McDonald and Punt, 2001). Thirty household groups are created, first disaggregated by population group African, Coloured, Indian and White (based on variable *headrace*)⁷, thereafter by location (urban and rural)⁸, and finally by income group using income percentiles. Urban and rural African households are split into quintiles (five groups). Urban Coloured households are divided into "triciles" (three groups), while rural Coloured households are split into "duociles" (two groups). Indian households are also split into "duociles". Finally, urban White households are disaggregated into quartiles (four groups), while rural White households are divided into "duociles". In all instances the high-income

⁶ The command can be interpreted as follows: sort the observations first by the unique household identification number (*hhid*) and then the unique person number (*s2persno*) of each individual in the household. Next, for each household, generate a variable *headrace* that is equal to the race of the first member *s3race[1]* – i.e. the head of the household – for each member of the household.

⁷ Note that McDonald and Punt (2001) used the IES 1995 racial classification rather than the OHS (*s3race/headrace*) classification and therefore did not take into account mixed-race households.

⁸ Indian households are not disaggregated by location since very few Indian households live in rural areas in South Africa.

group is split into two further groups. This further disaggregation is useful when looking at inequality in the distribution of income, as it is often the case that those households at the upper-end of the income scale earn a very large proportion of national income. Table 6 shows the household accounts.

Table 6: Household groups (SAM household accounts)

African		Coloured		Indian	White	
Urban	Rural	Urban	Rural	All	Urban	Rural
Quintile 1	Quintile 1	Tricile 1	Duocile 1	Duocile 1	Quartile 1	Duocile 1
Quintile 2	Quintile 2	Tricile 2	Duocile 2a	Duocile 2a	Quartile 2	Duocile 2a
Quintile 3	Quintile 3	Tricile 3a	Duocile 2b	Duocile 2b	Quartile 3	Duocile 2b
Quintile 4	Quintile 4	Tricile 3b	-	-	Quartile 4a	-
Quintile 5a	Quintile 5a	-	-	-	Quartile 4b	-
Quintile 6a	Quintile 6a	-	-	-	-	-

An important issue regarding the income group classification is the choice of measurement of income. Often total household income is used (*inctot*). However, this may introduce a degree of bias into the estimation. Using total income may lead to a situation where, for example, a household with 10 members earning a total income of R10000 per year is classified in the same group as another household with 1 member earning R10000. Clearly the latter is much better off and should be classified as being relatively richer than the first. One solution to this is using per capita income (*pcinc*). However, this also has its drawbacks. The size and structure of the household is important, as it is often the case that (1) large households benefit from economies of scale in the consumption of certain goods (e.g. housing), and (2) children require a lower level of consumption to satisfy basic needs. Consequently an adjusted per capita measure of income needs to be constructed, which adjusts the household income variable taking into account the size and structure of the household.

The adjusted household size variable E is constructed using the formula $E = (A + \alpha K)^\theta$, where α and θ are parameters that adjust for the lower living cost of children and economies of scale respectively. The variable A refers to the number of adults in a given household, while K refers to the number of children under the age of 10. May (1995, cited in Woolard and Leibbrandt, 1999) suggested setting $\alpha = 0.5$ and $\theta = 0.9$. For a more detailed discussion of the so-called equivalence scale approach see the forthcoming *Background Paper* on the formation of household groups.

The do-file **rhgroups.do**⁹ creates representative household based on total household income (see Table 6). Next, the total household income is adjusted for household size and structure in do-file **adjusted.do**. This do-file creates a variable *adinc*. Once this is done,

⁹ The name of the do-file is derived from representative household groups, the term used to describe these household groups.

rhgroupsad.do is performed, which classifies households as before in `rhgroups.do`, only now using the adjusted income variable. The user can select either household group classification variable. A comparison of the groups falls beyond the scope of this paper. The forthcoming *Background Paper* mentioned above will compare the two approaches in more detail.

4. Summary

The combined OHS-IES database contains important information not reflected in the individual datasets. The problems encountered when merging these sets were mentioned, and caution needs to be taken to ensure that information is credible. The technical appendix attached contains all the do-files that were used to construct the combined dataset. It also contains a number of (optional) do-files that were needed to organise data and create variables that can be used when exploring the dataset.

5. Technical Appendix

5.1. Variables contained in the OHS 1995 fixed-format files (layout.txt)

RECORD LAYOUT:

COVER PAGE AND SECTION 1 (HOUSEHOLDS)

COVER	(DISTRICT NUMBER)	(@1	3.)
COVER	(ENUMERATOR AREA NUMBER)	(@4	4.)
COVER	(VISITING POINT NUMBER)	(@8	2.)
COVER	(NUMBER OF HOUSEHOLDS AT VISITING POINT)	(@10	1.)
COVER	(TYPE OF ENUMERATION AREA (BRANCH OFFICE))	(@11	2.)
	SEE PAGE 3 OF QUEST. FOR CODES			
PAGE 3	(TYPE OF ENUMERATION AREA (ENUMERATOR))	(@13	2.)
Q1.1	(TYPE OF DWELLING1)	(@15	1.)
	(TYPE OF DWELLING2)	(@16	1.)
	(TYPE OF DWELLING3)	(@17	1.)
Q1.2	(MAIN TYPE OF DWELLING)	(@18	1.)
Q1.3	(OWNERSHIP)	(@19	1.)
Q1.4	(MAIN MATERIAL USED FOR ROOF)	(@20	2.)
	(MAIN MATERIAL USED FOR WALLS)	(@22	2.)
Q1.5	(ESTIMATED VALUE OF DWELLING)	(@24	7.)
Q1.6	(TOTAL NUMBER OF LIVINGROOMS (INCLUDE BEDR))	(@31	2.)
	(TOTAL NUMBER OF BEDROOMS)	(@33	2.)
Q1.7	(RAIN-WATER TANK)	(@35	1.)
Q1.8	(MAIN SOURCE OF DOMESTIC WATER (DRINKING))	(@36	2.)
	(MAIN SOURCE OF DOMESTIC WATER (OTHER))	(@38	2.)
Q1.9	(IS WATER OBTAINED ADEQUATE)	(@40	1.)
Q1.10	(DISTANCE TO WATER)	(@41	1.)
Q1.11	(DOES THE HOUSEHOLD PAY FOR THE WATER)	(@42	1.)
Q1.12	(COOKING - ELECTRICITY - PUBLIC SUPPLY)	(@43	1.)
	(COOKING - ELECTRICITY - GENERATOR)	(@44	1.)
	(COOKING - ELECTRICITY - SOLAR SYSTEM)	(@45	1.)
	(COOKING - GAS)	(@46	1.)
	(COOKING - PARAFFIN)	(@47	1.)
	(COOKING - WOOD)	(@48	1.)
	(COOKING - COAL)	(@49	1.)
	(COOKING - CHARCOAL)	(@50	1.)
	(COOKING - CROP WASTE)	(@51	1.)
	(COOKING - ANIMAL DUNG)	(@52	1.)
	(COOKING - OTHER)	(@53	1.)
	(HEATING - ELECTRICITY - PUBLIC SUPPLY)	(@54	1.)
	(HEATING - ELECTRICITY - GENERATOR)	(@55	1.)
	(HEATING - ELECTRICITY - BATTERY)	(@56	1.)
	(HEATING - ELECTRICITY - SOLAR SYSTEM)	(@57	1.)
	(HEATING - GAS)	(@58	1.)
	(HEATING - PARAFFIN)	(@59	1.)
	(HEATING - WOOD)	(@60	1.)
	(HEATING - COAL)	(@61	1.)
	(HEATING - CHARCOAL)	(@62	1.)
	(HEATING - CROP WASTE)	(@63	1.)
	(HEATING - ANIMAL DUNG)	(@64	1.)
	(HEATING - OTHER)	(@65	1.)
	(NO HEATING)	(@66	1.)
	(LIGHTING--ELECTRICITY - PUBLIC SUPPLY)	(@67	1.)
	(LIGHTING--ELECTRICITY - GENERATOR)	(@68	1.)
	(LIGHTING--ELECTRICITY - BATTERY)	(@69	1.)
	(LIGHTING--ELECTRICITY - SOLAR SYSTEM)	(@70	1.)
	(LIGHTING--GAS)	(@71	1.)
	(LIGHTING--PARAFFIN)	(@72	1.)
	(LIGHTING--CANDLE)	(@73	1.)
	(LIGHTING--OTHER)	(@74	1.)

Q1.13	(MAIN SOURCE OF FIRE WOOD)	(@75	1.)
Q1.14	(IS WOOD OBTAINED ADEQUATE)	(@76	1.)
Q1.15	(DISTANCE TO FETCH WOOD)	(@77	1.)
Q1.16	(DOES THE HOUSEHOLD PAY FOR THE WOOD)	(@78	1.)
Q1.17	(SANITATION IN DWELLING)	(@79	1.)
	(SANITATION ON SITE)	(@80	1.)
	(SANITATION OFF SITE)	(@81	1.)
Q1.18	(SANITATION - SHARE)	(@82	1.)
Q1.19	(SANITATION - SHARE - NO OF HOUSEHOLDS)	(@83	2.)
Q1.20	(HOW FAR IS THE NEAREST TOILET FACILITY)	(@85	1.)
Q1.21	(BUCKET TOILET - HOW FREQUENTLY REMOVED)	(@86	1.)
Q1.22	(REFUSE DISPOSAL)	(@87	1.)
Q1.23	(REFUSE REMOVED - HOW OFTEN)	(@88	1.)
Q1.24	(TELECOMMUNICATION - CELLULAR TELEPHONE)	(@89	1.)
	(TELECOMMUNICATION - OTHER TELEPHONE)	(@90	1.)
Q1.25	(DISTANCE TO NEAREST TELEPHONE)	(@91	1.)
Q1.26	(ACCESS TO LAND FOR FARMING)	(@92	1.)
	(PRIVATE LAND - TOTAL)	(@93	6.)
	(PRIVATE LAND - DRYLAND)	(@99	5.)
	(PRIVATE LAND - IRRIGATION - AREA)	(@104	5.)
	(PRIVATE LAND - IRRIGATION - SOURCE)	(@109	2.)
	(PRIVATE LAND - GRAZING)	(@111	1.)
	(COMMUNAL GARDEN- TOTAL)	(@112	6.)
	(COMMUNAL GARDEN- DRYLAND)	(@118	5.)
	(COMMUNAL GARDEN- IRRIGATION - AREA)	(@123	5.)
	(COMMUNAL GARDEN- IRRIGATION - SOURCE)	(@128	2.)
	(TRIBAL- COMMUNAL- GRAZING)	(@130	2.)
	(TRIBAL- RIGHT TO OCCUPY - TOTAL)	(@132	6.)
	(TRIBAL- RIGHT TO OCCUPY - DRYLAND)	(@138	5.)
	(TRIBAL- RIGHT TO OCCUPY - IRRIGATION - AREA)	(@143	5.)
	(TRIBAL- RIGHT TO OCCUPY - IRRIGATION - SRCE)	(@148	2.)
	(TRIBAL- COMMUNAL GARDEN - TOTAL)	(@150	6.)
	(TRIBAL- COMMUNAL GARDEN - DRYLAND)	(@156	5.)
	(TRIBAL- COMMUNAL GARDEN - IRRIGATION - AREA)	(@161	5.)
	(TRIBAL- COMMUNAL GARDEN - IRRIGATION - SRCE)	(@166	2.)
	(TRUST LAND - TOTAL)	(@168	6.)
	(TRUST LAND - DRYLAND)	(@174	5.)
	(TRUST LAND - IRRIGATION - AREA)	(@179	5.)
	(TRUST LAND - IRRIGATION - SOURCE)	(@184	2.)
	(TRUST LAND - GRAZING)	(@186	1.)
	(TENANCY LAND - TOTAL)	(@187	6.)
	(TENANCY LAND - DRYLAND)	(@193	5.)
	(TENANCY LAND - IRRIGATION - AREA)	(@198	5.)
	(TENANCY LAND - IRRIGATION - SOURCE)	(@203	2.)
	(TENANCY LAND - GRAZING)	(@205	1.)
Q1.27	(PHYSICAL SAFETY IN NEIGHBOURHOOD)	(@206	1.)
Q1.28	(PHYSICAL SAFETY IN DWELLING)	(@207	1.)
Q1.29	(PHYSICAL SAFETY CHANGED)	(@208	1.)
Q1.30	(VICTIM OF CRIME)	(@209	1.)
	(TYPE OF CRIME1)	(@210	1.)
	(TYPE OF CRIME2)	(@211	1.)
	(TYPE OF CRIME3)	(@212	1.)
	(TYPE OF CRIME4)	(@213	1.)
	(TYPE OF CRIME5)	(@214	1.)
	(TYPE OF CRIME6)	(@215	1.)
Q1.31	(SMOKE AND POLLUTION)	(@216	1.)
Q1.32	(MONEY TO FEED THE CHILDREN)	(@217	1.)
Q1.33	(LIVE THESE DAYS)	(@218	1.)
Q1.34	(COMPARE TO ONE YEAR AGO)	(@219	1.)
Q1.35	(RESIDING HEAD OF HOUSEHOLD)	(@220	1.)
	(REASON WHY HEAD DOES NOT RESIDE)	(@221	1.)
Q1.36	(STREET CHILDREN)	(@222	1.)
	(NUMBER OF STREET CHILDREN)	(@223	1.)
Q1.37	(SEEK MEDICAL HELP)	(@224	1.)
Q1.38	(DISTANCE TO MEDICAL HELP)	(@225	1.)
	(TIME OF JOURNEY TO MEDICAL HELP)	(@226	1.)

Q1.39	(DISTRICT OF MEDICAL HELP)	(@227	3.)
	(PROVINCE OF MEDICAL HELP)	(@230	1.)
Q1.40	(DISTRICT OF PUBLIC HOSPITAL)	(@231	3.)
	(PROVINCE OF PUBLIC HOSPITAL)	(@234	1.)
Q1.41	(DISTANCE TO WELFARE SERVICE)	(@235	1.)
BACK PAGE	(HOUSEHOLD1 - NUMBER OF PERSONS)	(@236	2.)
(MORE	(HOUSEHOLD2 - NUMBER OF PERSONS)	(@238	2.)
THAN ONE	(HOUSEHOLD3 - NUMBER OF PERSONS)	(@240	2.)
HOUSEHOLD	(HOUSEHOLD4 - NUMBER OF PERSONS)	(@242	2.)
AT VISIT-	(HOUSEHOLD5 - NUMBER OF PERSONS)	(@244	2.)
ING PNT	(HOUSEHOLD6 - NUMBER OF PERSONS)	(@246	2.)
	(PROVINCE NUMBER)	(@248	1.)
	(POPULATION GROUP)	(@249	1.)
	(HOUSEHOLD SIZE * NEW VARIABLE)	(@250	2.)
	(WEIGHT (TO BE DIVIDED BY 1000))	(@252	7.)

FLAP AND SECTION 2 (PERSONS)				
	(DISTRICT NUMBER)	(@1	3.)
	(ENUMERATOR AREA NUMBER)	(@4	4.)
	(VISITING POINT NUMBER)	(@8	2.)
FLAP	(PERSON)	(@10	2.)
B	(MEMBER PRESENT)	(@12	1.)
C	(GENDER)	(@13	1.)
D	(AGE)	(@14	2.)
D	(YEAR OF BIRTH)	(@16	4.)
E	(WOMEN WHO HAVE GIVEN BIRTH)	(@20	1.)
F	(IDEAL NUMBER OF CHILDREN)	(@21	1.)
Q2.1	(POPULATION GROUP)	(@22	1.)
Q2.2	(RELATIONSHIP)	(@23	1.)
Q2.3	(FATHER ALIVE)	(@24	1.)
Q2.3	(MOTHER ALIVE)	(@25	1.)
Q2.4	(FATHER NUMBER)	(@26	2.)
Q2.4	(MOTHER NUMBER)	(@28	2.)
Q2.5	(MARITAL STATUS)	(@30	1.)
Q2.6	(SPOUSE NUMBER)	(@31	2.)
Q2.7	(PROVINCE BORN)	(@33	1.)
	(DISTRICT BORN)	(@34	3.)
	(COUNTRY BORN)	(@37	2.)
Q2.8	(MOVE INTO THIS AREA?)	(@39	1.)
	(MIGRATION PROVINCE)	(@40	1.)
	(MIGRATION DISTRICT)	(@41	3.)
	(MIGRATION COUNTRY)	(@44	2.)
Q2.9	(TYPE OF PREVIOUS DWELLING)	(@46	1.)
Q2.10	(ATTENDING PRE-SCHOOL, CRECHE)	(@47	1.)
Q2.11	(SCHOLAR/STUDENT?)	(@48	1.)
Q2.12	(ATTEND LITERACY PROGRAM)	(@49	1.)
Q2.13	(ACQUIRE OPERATOR SKILL)	(@50	1.)
	(TYPE OF SKILL)	(@51	3.)
Q2.14	(MIGRANT WORKER)	(@54	1.)
Q2.15	(SCHOOL FEEDING SCHEME)	(@55	1.)
Q2.16	(HIGHEST LEVEL OF EDUCATION)	(@56	2.)
Q2.17	(WISH TO CONTINUE EDUCATION)	(@58	1.)
	(REASON FOR NOT CONTINUEING)	(@59	1.)
Q2.18	(SCHOOL FEES)	(@60	5.)
	(SCHOOL TRANSPORT FEES)	(@65	5.)
	(OTHER SCHOOL FEES)	(@70	5.)
Q2.19	(ILLNESS)	(@75	1.)
Q2.20	(INJURY)	(@76	1.)
Q2.21	(DISCHARGED FROM HOSPITAL)	(@77	1.)
Q2.22	(CONSULT NURSE)	(@78	1.)
	(CONSULT SPECIALIST)	(@79	1.)
	(CONSULT DOCTOR)	(@80	1.)
	(CONSULT DENTIST)	(@81	1.)
	(CONSULT PHARMACIST)	(@82	1.)
	(CONSULT OTHER MEDICAL)	(@83	1.)

	(CONSULT FAITH HEALER)	(@84	1.)
	(CONSULT SANGOMA)	(@85	1.)
Q2.23	(WHERE DID CONSULTATION TAKE PLACE)	(@86	1.)
Q2.24	(DID THE HOUSEHOLD HAVE TO PAY)	(@87	1.)
Q2.25	(ACCESS TO MEDICAL AID BENEFIT FUND)	(@88	1.)
Q2.26	(DOES ... SMOKE?)	(@89	1.)
	(AGE STARTED SMOKING)	(@90	2.)
Q2.27	(DISABILITY1)	(@92	1.)
	(DISABILITY2)	(@93	1.)
	(DISABILITY3)	(@94	1.)
	(DISABILITY4)	(@95	1.)
Q2.28	(MAKE USE OF SOCIAL WELFARE SERVICE)	(@96	1.)
	(WELFARE SERVICE1)	(@97	1.)
	(WELFARE SERVICE2)	(@98	1.)
	(WELFARE SERVICE3)	(@99	1.)
	(PROVINCE NUMBER)	(@100	1.)
	(TYPE OF ENUMERATION AREA (BRANCH OFFICE))	(@101	2.)
	SEE PAGE 3 OF QUEST. FOR CODES			
	(MAIN DWELLING - SECTION 1 USE Q1.1 AND Q1.2))	(@103	1.)
	(WEIGHT (TO BE DIVIDED BY 1000))	(@104	7.)

SECTION 3 (WORKERS)				
	(DISTRICT NUMBER)	(@1	3.)
	(ENUMARATOR AREA NUMBER)	(@4	4.)
	(VISITING POINT NUMBER)	(@8	2.)
	(PERSON NUMBER)	(@10	2.)
Q3.1	(ACTIVITIES LAST 7 DAYS)	(@12	2.)
Q3.2	(ANY WORK DURING PAST YEAR)	(@14	1.)
	(ANY WORK - HOW LONG AGO?)	(@15	1.)
Q3.3	(HOURS WORKED)	(@16	2.)
Q3.4	(WOULD ... HAVE LIKED TO WORK MORE HOURS?)	(@18	1.)
	(HOURS IN TOTAL)	(@19	2.)
Q3.5	(TRANSPORT1)	(@21	1.)
	(TRANSPORT2)	(@22	1.)
	(TRANSPORT3)	(@23	1.)
Q3.6	(TIME LEAVE HOME FOR WORK)	(@24	4.)
Q3.7	(TIME ARRIVE AT WORK)	(@28	4.)
Q3.8	(WORK DISTRICT)	(@32	3.)
	(WORK PROVINCE)	(@35	1.)
Q3.9	(WORKSTATUS)	(@36	1.)
Q3.12	(YEAR AND MONTH STARTED WORKING)	(@37	4.)
Q3.13	(MEMBER OF TRADE UNION)	(@41	1.)
Q3.14	(ENTITLED TO MATERNITY LEAVE)	(@42	1.)
Q3.15	(OCCUPATION OF EMPLOYEE)	(@43	3.)
Q3.11	(INDUSTRY OF EMPLOYEE)	(@46	2.)
Q3.16	(TOTAL SALARY/PAY (RAND))	(@48	7.)
	(TOTAL SALARY PAY (INCOME GROUP))	(@55	2.)
	(INCOME INTERVAL)	(@57	1.)
	(NATURA - TRANSPORT)	(@58	4.)
	(NATURA - FOOD)	(@62	4.)
	(NATURA - OTHER)	(@66	4.)
Q3.17	(DEDUCTIONS (RAND))	(@70	5.)
	(DEDUCTIONS INTERVAL)	(@75	1.)
Q3.18	(ADDITIONAL WORK)	(@76	1.)
Q3.19	(OCCUPATION OF EMPLOYER/OWN ACCOUNT WORKER)	(@77	3.)
	(INDUSTRY OF EMPLOYER/OWN ACCOUNT WORKER)	(@80	2.)
Q3.20	(REGISTRATION)	(@82	1.)
Q3.21	(VAT)	(@83	1.)
Q3.22	(GROSS INCOME OF EMPLOYER (RAND))	(@84	7.)
	(GROSS INCOME OF EMPLOYER (CODE))	(@91	2.)
	(INCOME INTERVAL)	(@93	1.)
Q3.23	(EXPENSES - SALARIES, COMMISSION, GOODS)	(@94	5.)
Q3.24	(TOTAL UNPAID EMPLOYEES)	(@99	2.)
	(UNPAID EMPLOYEES - UNDER 15 YEARS)	(@101	2.)
	(TOTAL PAID EMPLOYEES)	(@103	2.)

	(PAID EMPLOYEES - UNDER 15 YEARS)	(@105	2.)
	(EXPENSES - SALARIES)	(@107	5.)
	(FAMILY WORKERS1)	(@112	2.)
	(FAMILY WORKERS2)	(@114	2.)
Q3.25	(WAS WORK DONE DURING LAST 7 DAYS?)	(@116	1.)
Q3.26	(JOB ATTACHMENT)	(@117	1.)
Q3.27	(REASON FOR ABSENCE FROM WORK)	(@118	2.)
Q3.28	(SUITABLE JOB OFFERED - WILL YOU ACCEPT IT)	(@120	1.)
Q3.29	(HOW LONG BEEN SEEKING WORK?)	(@121	1.)
Q3.30	(METHOD1 WORK SOUGHT)	(@122	1.)
	(METHOD2 WORK SOUGHT)	(@123	1.)
	(METHOD3 WORK SOUGHT)	(@124	1.)
Q3.31	(HAS UNEMPLOYED WORKED BEFORE)	(@125	1.)
Q3.32	(PREVIOUS OCCUPATION OF UNEMPLOYED)	(@126	3.)
Q3.33	(REASON NOT WORKING LAST 7 DAYS)	(@129	2.)
Q3.34	(SUPPORT1)	(@131	1.)
	(SUPPORT2)	(@132	1.)
	(SUPPORT3)	(@133	1.)
	(PROVINCE NUMBER)	(@134	1.)
	(TYPE OF ENUMERATION AREA (BRANCH OFFICE))	(@135	2.)
	SEE PAGE 3 OF QUEST. FOR CODES			
	(AGE)	(@137	2.)
	(GENDER)	(@139	1.)
	(POPULATION GROUP)	(@140	1.)
	(HIGHEST LEVEL OF EDUCATION)	(@141	2.)

NEW CODES CREATED

	(WORKERS (15 YEARS AND OLDER) (CODES 1 AND 2))	(@143	1.)
	1= WORKED PAST 7 DAYS, 2= HAS ATTACHMENT TO A JOB			
	(EXPANDED UNEMPLOYED =1)	(@144	1.)
	(STRICT UNEMPLOYED =1)	(@145	1.)
	(NOT ECON ACTIVE (15 YEARS AND OLDER)=1)	(@146	1.)
	(OCCUPATION -MAIN GROUPS)	(@147	3.)
	(INDUSTRY - MAIN GROUPS)	(@150	2.)
	(TIME TO GET TO WORK - USE Q3.6 AND Q3.7)	(@152	3.)
	(INCOME OF EMPLOYEE (RAND) CALC /MONTH)	(@155	7.)
	(INCOME OF EMPLOYEE (CODE) CALC /MONTH)	(@162	2.)
	(INCOME OF EMPLOYER (RAND) CALC /MONTH)	(@164	7.)
	(INCOME OF EMPLOYER (CODE) CALC /MONTH)	(@171	2.)
	(WEIGHT (DIVIDE BY 1000))	(@173	7.)

5.2. Listing of do-files used to create the combined dataset

5.2.1. *combined.do*

```
#delimit;

*This do-file creates a database combined.dta*;
*First reads in OHS data to form ohs95.dta*;
*Next merges ohs95.dta with ies95original.dta*;
*Finally creates representative household groups*;

set more off;
do readohs.do;

do merge.do;

household.do;

save combined.dta, replace;

*=====*;
*           THE END           *;
*=====*;
```

5.2.2. *merge.do*

```
#delimit;

set more off;

clear;

*Opening house.dta and keeping required variables*;
*=====*;

use house.dta, clear ;
keep hhid slhhsz ;
sort hhid;
save houseshort.dta, replace;

*Opening person.dta and keeping required variables*;
*=====*;

use person.dta, clear ;
keep hhid s2persno s2age q2_16;
do education.do;
do agegroup.do;
sort hhid;
save personshort.dta, replace;

*Opening work.dta and keeping required variables*;
*Also perform occcodes.do and income.do*;
*=====*;

use work.dta, clear ;
do occcodes.do;
do income.do;
keep hhid q3_1 q3_11 q3_15 q3_16* q3_22* s3gender s3new1 s3new2
      s3race occcodes factors incself incoth incempl;
sort hhid;
save workshort.dta, replace;
```

```

*Merging OHS datasets*;
*=====*;

merge hhid using houseshort.dta;
    rename _merge merge1;
    sort hhid;
merge hhid using personsshort.dta;
    rename _merge merge2;
    sort hhid;
do structure.do;
save ohs95.dta, replace;

*Opening IES dataset copied from IES folder and renamed ies95original.dta*;
*Also create household groups after merge with OHS dataset*;
*=====*;

use ies95original.dta, clear;
keep hhid _problem race settle extot inc* p*;
rename race iesrace;

save ies95short.dta, replace;
    sort hhid;
merge hhid using ohs95.dta;
    rename _merge merge3;
    tab merge1;
    tab merge2;
    tab merge3;

keep if merge3==3;
drop if _problem == 1 | _problem == 2 | _problem == 3;

save combined.dta, replace;

```

5.2.3. *education.do*

```

#delimit;

*education variable*;

gen educ      = 0 if q2_16 == 0;
replace educ = 1 if q2_16 >= 1 & q2_16 <= 5;
replace educ = 2 if q2_16 >= 6 & q2_16 <= 8;
replace educ = 3 if q2_16 >= 9 & q2_16 <= 10;
replace educ = 4 if q2_16 >= 11 & q2_16 <= 13;

label var educ "Educational attainment";
label define education 0 "No schooling" 1 "Primary school" 2 "Std 6, 7 or
8"
                3 "Std 9 or 10" 4 "Tertiary qualification";
label values educ education;

```

5.2.4. *agegroup.do*

```

#delimit;
set more off;

gen agegroup=1 if s2age >= 10 & s2age <= 14;
replace agegroup=2 if s2age >= 15 & s2age <= 19;
replace agegroup=3 if s2age >= 20 & s2age <= 24;
replace agegroup=4 if s2age >= 25 & s2age <= 29;
replace agegroup=5 if s2age >= 30 & s2age <= 34;

```

```

replace agegroup=6 if s2age >= 35 & s2age <= 39;
replace agegroup=7 if s2age >= 40 & s2age <= 44;
replace agegroup=8 if s2age >= 45 & s2age <= 49;
replace agegroup=9 if s2age >= 50 & s2age <= 54;
replace agegroup=10 if s2age >= 55 & s2age <= 59;
replace agegroup=11 if s2age >= 60 & s2age <= 64;
replace agegroup=12 if s2age >= 65 & s2age <= 200;

label var agegroup "Age groups/categories";
label define agegr      1 "Age 10-14" 2 "Age 15-19" 3 "Age 20-24"
                      4 "Age 25-29" 5 "Age 30-34" 6 "Age 35-39"
                      7 "Age 40-44" 8 "Age 45-49" 9 "Age 50-54"
                      10 "Age 55-59" 11 "Age 60-64" 12 "Age 65+";
label values agegroup agegr;

```

5.2.5. *occcodes.do*

```

#delimit;

set more off;

gen origocc = q3_15;

replace q3_15 = 10 if q3_15 == 934;
replace q3_15 = 1 if q3_15 >= 100 & q3_15 <= 199;
replace q3_15 = 2 if q3_15 >= 200 & q3_15 <= 299;
replace q3_15 = 3 if q3_15 >= 300 & q3_15 <= 399;
replace q3_15 = 4 if q3_15 >= 400 & q3_15 <= 499;
replace q3_15 = 5 if q3_15 >= 500 & q3_15 <= 599;
replace q3_15 = 6 if q3_15 >= 600 & q3_15 <= 699;
replace q3_15 = 7 if q3_15 >= 700 & q3_15 <= 799;
replace q3_15 = 8 if q3_15 >= 800 & q3_15 <= 899;
replace q3_15 = 9 if q3_15 >= 900 & q3_15 <= 999;

gen occcodes = q3_15 + 100 if s3race == 1;
replace occcodes = q3_15 + 200 if s3race == 2;
replace occcodes = q3_15 + 300 if s3race == 3;
replace occcodes = q3_15 + 400 if s3race == 4;

gen factors = 2 if occcodes == 101;
replace factors = 3 if occcodes == 102;
replace factors = 4 if occcodes == 103;
replace factors = 5 if occcodes == 104;
replace factors = 6 if occcodes == 105;
replace factors = 7 if occcodes == 106;
replace factors = 8 if occcodes == 107;
replace factors = 9 if occcodes == 108;
replace factors = 10 if occcodes == 109;
replace factors = 11 if occcodes == 110;
replace factors = 12 if occcodes == 111;
replace factors = 13 if occcodes == 201;
replace factors = 14 if occcodes == 202;
replace factors = 15 if occcodes == 203;
replace factors = 16 if occcodes == 204;
replace factors = 17 if occcodes == 205;
replace factors = 18 if occcodes == 206;
replace factors = 19 if occcodes == 207;
replace factors = 20 if occcodes == 208;
replace factors = 21 if occcodes == 209;
replace factors = 22 if occcodes == 210;
replace factors = 23 if occcodes == 211;
replace factors = 24 if occcodes == 301;
replace factors = 25 if occcodes == 302;
replace factors = 26 if occcodes == 303;
replace factors = 27 if occcodes == 304;

```



```

replace factors = 28 if occcodes == 305;
replace factors = 29 if occcodes == 306;
replace factors = 30 if occcodes == 307;
replace factors = 31 if occcodes == 308;
replace factors = 32 if occcodes == 309;
replace factors = 33 if occcodes == 310;
replace factors = 34 if occcodes == 311;
replace factors = 35 if occcodes == 401;
replace factors = 36 if occcodes == 402;
replace factors = 37 if occcodes == 403;
replace factors = 38 if occcodes == 404;
replace factors = 39 if occcodes == 405;
replace factors = 40 if occcodes == 406;
replace factors = 41 if occcodes == 407;
replace factors = 42 if occcodes == 408;
replace factors = 43 if occcodes == 409;
replace factors = 44 if occcodes == 410;
replace factors = 45 if occcodes == 411;

replace factors = 91 if occcodes == 100;
replace factors = 92 if occcodes == 200;
replace factors = 93 if occcodes == 300;
replace factors = 94 if occcodes == 400;

```

5.2.6. *income.do*

```

#delimit;
set more off;

*Income from main job (working for someone else);
*=====;

gen incb = 0;
replace incb = 0 if q3_16b == 1;
replace incb = 500 if q3_16b == 2;
replace incb = 1125 if q3_16b == 3;
replace incb = 1375 if q3_16b == 4;
replace incb = 1750 if q3_16b == 5;
replace incb = 2250 if q3_16b == 6;
replace incb = 2750 if q3_16b == 7;
replace incb = 3500 if q3_16b == 8;
replace incb = 5000 if q3_16b == 9;
replace incb = 7000 if q3_16b == 10;
replace incb = 9000 if q3_16b == 11;
replace incb = 11250 if q3_16b == 12;
replace incb = 13750 if q3_16b == 13;
replace incb = 17500 if q3_16b == 14;
replace incb = 22500 if q3_16b == 15;
replace incb = 27500 if q3_16b == 16;
replace incb = 35000 if q3_16b == 17;
replace incb = 50000 if q3_16b == 18;
replace incb = 70000 if q3_16b == 19;
replace incb = 90000 if q3_16b == 20;
replace incb = 112500 if q3_16b == 21;
replace incb = 137500 if q3_16b == 22;
replace incb = 175000 if q3_16b == 23;
replace incb = 225000 if q3_16b == 24;
replace incb = 275000 if q3_16b == 25;
replace incb = 350000 if q3_16b == 26;
replace incb = 450000 if q3_16b == 27;
replace incb = 550000 if q3_16b == 28;
replace incb = 750000 if q3_16b == 29;
replace incb = 0 if q3_16b == 30;

gen inca = q3_16a ;

```

```

replace inca = inca * 240 if q3_16c == 1;
*Assumption: 240 out of 365 days worked*;
replace inca = inca * 52 if q3_16c == 2;
replace inca = inca * 12 if q3_16c == 3;
replace inca = inca * 1 if q3_16c == 4;
replace inca = 0 if inca == 9999999;

replace incb = incb * 240 if q3_16c == 1;
*Assumption: 240 out of 365 days worked*;
replace incb = incb * 52 if q3_16c == 2;
replace incb = incb * 12 if q3_16c == 3;
replace incb = incb * 1 if q3_16c == 4;

gen incoth= inca+incb;
label var incoth "Total income from main job 3.16";

*Income from working for oneself ;
*=====;

gen incd = 0;
replace incd = 0 if q3_22b == 1;
replace incd = 500 if q3_22b == 2;
replace incd = 1125 if q3_22b == 3;
replace incd = 1375 if q3_22b == 4;
replace incd = 1750 if q3_22b == 5;
replace incd = 2250 if q3_22b == 6;
replace incd = 2750 if q3_22b == 7;
replace incd = 3500 if q3_22b == 8;
replace incd = 5000 if q3_22b == 9;
replace incd = 7000 if q3_22b == 10;
replace incd = 9000 if q3_22b == 11;
replace incd = 11250 if q3_22b == 12;
replace incd = 13750 if q3_22b == 13;
replace incd = 17500 if q3_22b == 14;
replace incd = 22500 if q3_22b == 15;
replace incd = 27500 if q3_22b == 16;
replace incd = 35000 if q3_22b == 17;
replace incd = 50000 if q3_22b == 18;
replace incd = 70000 if q3_22b == 19;
replace incd = 90000 if q3_22b == 20;
replace incd = 112500 if q3_22b == 21;
replace incd = 137500 if q3_22b == 22;
replace incd = 175000 if q3_22b == 23;
replace incd = 225000 if q3_22b == 24;
replace incd = 275000 if q3_22b == 25;
replace incd = 350000 if q3_22b == 26;
replace incd = 450000 if q3_22b == 27;
replace incd = 550000 if q3_22b == 28;
replace incd = 750000 if q3_22b == 29;
replace incd = 0 if q3_22b == 30;

gen incc = q3_22a ;
replace incc = incc * 240 if q3_22c == 1;
*Assumption: 240 out of 365 days worked*;
replace incc = incc * 52 if q3_22c == 2;
replace incc = incc * 12 if q3_22c == 3;
replace incc = incc * 1 if q3_22c == 4;
replace incc = 0 if incc == 9999999;

replace incd = incd * 240 if q3_22c == 1;
*Assumption: 240 out of 365 days worked*;
replace incd = incd * 52 if q3_22c == 2;
replace incd = incd * 12 if q3_22c == 3;
replace incd = incd * 1 if q3_22c == 4;

```

```

gen incself= incc+incd;
label var incself "Total income from self employment 3.22";

gen incempl = incself + incoth;
label var incempl "Total income from employment - self and other";

drop inca incb incc incd;

```

5.2.7. *structure.do*

```

#delimit;
set more off;

gen structure = .;
replace structure = 1 if (s3new1 == 1 | s3new1 == 2) & s2age>=15 ;
replace structure = 2 if (q3_1 == 4) & s2age>=15 ;
replace structure = 3 if (q3_1 == 8 | q3_1 == 9) & s2age>=15 ;
replace structure = 4 if (s3new2==1) & s2age>=15 ;
replace structure = 6 if s2age<15 ;
replace structure = 5 if structure == . ;

label var structure "Household structure";
label define struct 1 "Working full-/part-time 15+" 2 "Students 15+"
3 "Retired/unable to work 15+" 4 "Unemployed expanded 15+" 5 "Other 15+" 6
"Under age 0-15";
label values structure struct;

```

5.2.8. *households.do*

```

#delimit;

*Forms representative household groups using *;
*   (1) household income*;
*   (2) adult equivalent household income*;
*Then generates a variable per capita income*;

set more off;
do rhgroups.do;
save combined.dta, replace;

do adjusted.do;
do rhgroupsad.do;
sort hhid;
save combine2.dta, replace;

use combined.dta, clear;
merge hhid using combine2.dta;

keep hhid settle extot inctot inclab incgos inctrans inccorp agegroup edu*
q2_16
    incgov incother occcodes factors s3race incoth incself adinc q3_15
s3gender
    incempl slhhsz s2persno s2age A K headrace iesrace hhgroup hhgrad
q3_11 ;

gen pcinc = inctot/slhhsz;

```

5.2.9. *rhgroups.do*

(Not included here. This do-file is similar to *rhgroupsad.do*, the only difference being the use of *inctot* instead of *adinc*.)

5.2.10. *adjusted.do*

```
#delimit;

gen adults = 1 if s2age > 10;
keep if adults == 1;
sort hhid;
by hhid: gen A = _N;
gen K = slhhsz-A;
keep if s2persno==1;

*assume: alpha = 0.5 and theta = 0.9 ;

gen E=(A+(0.5*K))^0.9;
gen adinc = inctot/E;
gen pcinc = inctot/slhhsz;
```

5.2.11. *rhgroupsad.do*

```
*Same as rhgroups but income adjusted - adult equivalent*
*=====*
```

Note that the data is individual level. The merge had to be done first
 * because we use the OHS race variable for consistency reasons. Since*
 * household groups have to be at household level income to avoid double*
 * counting of income, the headrace variable is first created for only the*
 * of the household, and thereafter extended to the rest of set for
 analysis*;

```
#delimit;
set more off;

drop  auperc auquint arperc arquint cuperc cutric crperc
      crduo aperc aduo wuperc wuquart wrperc wrduo;

*INCOME QUINTILES: URBAN AFRICANS*
*=====*
```

```
pctile auperc = adinc if headrace == 1 & settle == 1, nq(10);
egen maxadinc = max(adinc);
sort auperc;
generate auquint = .;
replace auquint = 1 if adinc >= 0 & adinc <= auperc[2] & headrace == 1 &
settle == 1;
replace auquint = 2 if adinc > auperc[2] & adinc <= auperc[4] & headrace ==
1 & settle == 1;
replace auquint = 3 if adinc > auperc[4] & adinc <= auperc[6] & headrace ==
1 & settle == 1;
replace auquint = 4 if adinc > auperc[6] & adinc <= auperc[8] & headrace ==
1 & settle == 1;
replace auquint = 5 if adinc > auperc[8] & adinc <= auperc[9] & headrace ==
1 & settle == 1;
replace auquint = 6 if adinc > auperc[9] & adinc <= maxadinc & headrace ==
1 & settle == 1;

label values auquint quintlab;
label var auquint "African urban quintiles";

*INCOME QUINTILES: RURAL AFRICANS*
;
pctile arperc = adinc if headrace == 1 & settle == 2, nq(10);
```

```

sort arperc;
generate arquint = .;
replace arquint = 1 if adinc >= 0 & adinc <= arperc[2] & headrace == 1 &
settle == 2;
replace arquint = 2 if adinc > arperc[2] & adinc <= arperc[4] & headrace ==
1 & settle == 2;
replace arquint = 3 if adinc > arperc[4] & adinc <= arperc[6] & headrace ==
1 & settle == 2;
replace arquint = 4 if adinc > arperc[6] & adinc <= arperc[8] & headrace ==
1 & settle == 2;
replace arquint = 5 if adinc > arperc[8] & adinc <= arperc[9] & headrace ==
1 & settle == 2;
replace arquint = 6 if adinc > arperc[9] & adinc <= maxadinc & headrace ==
1 & settle == 2;

label values arquint quintlab;
label var arquint "African rural quintiles";

*INCOME TRICILES: URBAN COLOUREDS*
;
pctile cuperc = adinc if headrace == 2 & settle == 1, nq(6);
sort cuperc;
generate cutric = .;
replace cutric = 1 if adinc >= 0 & adinc <= cuperc[2] & headrace == 2 &
settle == 1;
replace cutric = 2 if adinc > cuperc[2] & adinc <= cuperc[4] & headrace ==
2 & settle == 1;
replace cutric = 3 if adinc > cuperc[4] & adinc <= cuperc[5] & headrace ==
2 & settle == 1;
replace cutric = 4 if adinc > cuperc[5] & adinc <= maxadinc & headrace == 2
& settle == 1;

label values cutric triclab;
label var cutric "Coloured urban triciles";

*INCOME DUOCILES: RURAL COLOUREDS*
;
pctile crperc = adinc if headrace == 2 & settle == 2, nq(4);
sort crperc;
generate crduo = .;
replace crduo = 1 if adinc >= 0 & adinc <= crperc[2] & headrace == 2 &
settle == 2;
replace crduo = 2 if adinc > crperc[2] & adinc <= crperc[3] & headrace == 2
& settle == 2;
replace crduo = 3 if adinc > crperc[3] & adinc <= maxadinc & headrace == 2
& settle == 2;

label values crduo duolab;
label var crduo "Coloured rural duociles";

*INCOME DUOCILES: ALL INDIANS*
;
pctile aperc = adinc if headrace == 3, nq(4);
sort aperc;
generate aduo = .;
replace aduo = 1 if adinc >= 0 & adinc <= aperc[2] & headrace == 3;
replace aduo = 2 if adinc > aperc[2] & adinc <= aperc[3] & headrace == 3;
replace aduo = 3 if adinc > aperc[3] & adinc <= maxadinc & headrace == 3;

label values aduo duolab;
label var aduo "Indian duociles";

*INCOME QUARTILES: URBAN WHITES*
;
pctile wuperc = adinc if headrace == 4 & settle == 1, nq(8);
sort wuperc;

```

```

generate wuquart = .;
replace wuquart = 1 if adinc >= 0 & adinc <= wuperc[2] & headrace == 4 &
settle == 1;
replace wuquart = 2 if adinc > wuperc[2] & adinc <= wuperc[4] & headrace ==
4 & settle == 1;
replace wuquart = 3 if adinc > wuperc[4] & adinc <= wuperc[6] & headrace ==
4 & settle == 1;
replace wuquart = 4 if adinc > wuperc[6] & adinc <= wuperc[7] & headrace ==
4 & settle == 1;
replace wuquart = 5 if adinc > wuperc[7] & adinc <= maxadinc & headrace ==
4 & settle == 1;

label values wuquart quartlab;
label var wuquart "White urban quartiles";

*INCOME DUOCILES: RURAL WHITES*
;
pctile wrperc = adinc if headrace == 4 & settle == 2, nq(4);
sort wrperc;
generate wrduo = .;
replace wrduo = 1 if adinc >= 0 & adinc <= wrperc[2] & headrace == 4 &
settle == 2;
replace wrduo = 2 if adinc > wrperc[2] & adinc <= wrperc[3] & headrace == 4
& settle == 2;
replace wrduo = 3 if adinc > wrperc[3] & adinc <= maxadinc & headrace == 4
& settle == 2;

label values wrduo duolab;
label var wrduo "White rural duociles";

tab auquint;
tab arquint;
tab cutric;
tab crduo;
tab aduo;
tab wuquart;
tab wrduo;

*Creating single hhgrad variable at individual level*;
*=====*;

sort hhid s2persno;
by hhid: generate hhgrad = 1 if auquint[1] == 1;
by hhid: replace hhgrad = 2 if auquint[1] == 2;
by hhid: replace hhgrad = 3 if auquint[1] == 3;
by hhid: replace hhgrad = 4 if auquint[1] == 4;
by hhid: replace hhgrad = 5 if auquint[1] == 5;
by hhid: replace hhgrad = 6 if auquint[1] == 6;

by hhid: replace hhgrad = 7 if arquint[1] == 1;
by hhid: replace hhgrad = 8 if arquint[1] == 2;
by hhid: replace hhgrad = 9 if arquint[1] == 3;
by hhid: replace hhgrad = 10 if arquint[1] == 4;
by hhid: replace hhgrad = 11 if arquint[1] == 5;
by hhid: replace hhgrad = 12 if arquint[1] == 6;

by hhid: replace hhgrad = 13 if cutric[1] == 1;
by hhid: replace hhgrad = 14 if cutric[1] == 2;
by hhid: replace hhgrad = 15 if cutric[1] == 3;
by hhid: replace hhgrad = 16 if cutric[1] == 4;

by hhid: replace hhgrad = 17 if crduo[1] == 1;
by hhid: replace hhgrad = 18 if crduo[1] == 2;
by hhid: replace hhgrad = 19 if crduo[1] == 3;

by hhid: replace hhgrad = 20 if aduo[1] == 1;

```

```
by hhid: replace hhgrad = 21 if aduo[1] == 2;
by hhid: replace hhgrad = 22 if aduo[1] == 3;

by hhid: replace hhgrad = 23 if wuquart[1] == 1;
by hhid: replace hhgrad = 24 if wuquart[1] == 2;
by hhid: replace hhgrad = 25 if wuquart[1] == 3;
by hhid: replace hhgrad = 26 if wuquart[1] == 4;
by hhid: replace hhgrad = 27 if wuquart[1] == 5;

by hhid: replace hhgrad = 28 if wrduo[1] == 1;
by hhid: replace hhgrad = 29 if wrduo[1] == 2;
by hhid: replace hhgrad = 30 if wrduo[1] == 3;

label var hhgrad "RHG: Adult Equivalent Income";

label values hhgrad grouplab;
```

Technical Papers in this Series

Number	Title	Date
TP2003: 1	Creating a 1995 IES Database in STATA	September 2003
TP2003: 2	Creating a 1995 OHS and a combined OHS-IES Database in STATA	September 2003
TP2003: 3	A Standard Computable General Equilibrium Model Version 3: Technical Documentation	September 2003
TP2003: 4	Measures of Poverty and Inequality: A Reference paper	October 2003
TP2004: 1	SeeResults: A spreadsheet Application for the Analysis of CGE Model Results	November 2004
TP2004: 2	The Organisation of Trade Data for inclusion in Social Accounting Matrix	December 2004
TP2004: 3	Creating a 2000 IES Database in Stata	August 2004
TP2004: 4	Forming Representative Household Groups in a SAM	July 2004

Other PROVIDE Publications

Background Paper Series

Working Papers

Research Reports