The following text is largely an excerpt from:

South Africa Labour Development Research Unit. 1994. "South Africans Rich and Poor: Baseline Household Statistics." University of Cape Town, South Africa.

OVERVIEW OF THE SOUTH AFRICA INTEGRATED HOUSEHOLD SURVEY

INTRODUCTION

This note provides an overview of the South Africa Integrated Household Survey, which covered approximately 9000 households, drawn from a carefully selected sample throughout the length and breadth of South Africa. The principal purpose of the survey, which was undertaken during the nine months leading up to the countries first democratic elections at the end of April 1994, was to collect hard statistical information about the conditions under which South Africans live in order to provide policy makers with the data required for planning strategies to implement such goals as those outlined in the Government of National Unity's Reconstruction and Development Programme.

The idea for such a survey was first mooted by a delegation of South Africans, from the African National Congress and the Congress of South African Trade Unions, led by Mr. Thabo Mbeki, when they met officials of the International Bank for Reconstruction and Development (World Bank) in Washington in April 1992. Responding to the South African request for more thinking about effective strategies to combat poverty, the Word Bank sent a task force, led by Ms. Neeta Sirur, to the country to assess what needed to be done. As a result of this visit the Southern Africa Labour Development Research Unit (Saldru) in the School of Economics at the University of Cape Town was asked to coordinate and manage the collection of data required. In order to broaden the base of the process a small steering committee drawn from social scientists at all three universities in the Western Cape was appointed to oversee the project. At the same time a reference group of persons, drawn from across the political spectrum, was established in order to ensure that the process was as technically sound, politically legitimate, and ideologically unbiased as possible.

Funding for the Project was generously provided by the governments of Denmark, the Netherlands and Norway working through the World Bank whose participation in the Project enabled the South African team to draw on a wide range of international experience and advice. A notable feature of the process has been the fruitful interaction between South Africans responsible for the Survey and the staff and consultants of the Bank. What began as a debate between South Africans and officials of the World Bank about the Survey rapidly became a discussion amongst interested colleagues on how best to deal with the different problems (e.g. sampling) that emerged during the course of the Project. The model of a project of this nature, run by citizens of the country concerned in such a way as to enable creative inputs and interaction from and with an institution such as the World Bank, is, we believe, one that needs to be developed further.

Two important understandings were reached and agreed upon in the early negotiations. One was that the data obtained as a result of the Survey would be public property, available to anybody wished to make use of it. It would not belong to any particular research institute, university, government department, nor to the World Bank. The data, it was agreed would be placed in the public domain. In this way those involved in the Survey hoped to encourage and consolidate an attitude in South Africa that sees the public accessibility to all such data, from whatever source, as a fundamental attribute of a democratic society. It is in this spirit that the data files, documentation, and questionnaires are provided on the World Wide Web for anyone to download and use.

The second important understanding lay in recognition of the fact that collection of data was not the only goal. No less important was the need to ensure that the actual process of running the Project happened in such a way as to enlarge and strengthen the South African capacity to generate and to analyze such data. It was agreed that it was especially important to find ways of drawing upon the insights and experience, whilst simultaneously enhancing, the skills of South Africans in order to help overcome the legacies of Apartheid.

In order that the Survey might not take place in a vacuum, the World Bank suggested that a comprehensive search of the available literature be undertaken in order to collate all information about living standards and development in South Africa just prior to the start of the Survey itself. Basically this was an attempt to document how the situation had or had not changed in the decade since the main empirical work was done for the Second Carnegie Inquiry Into Poverty & Development in 1983/84. Social scientists were drawn in from universities and other research organizations around the country, workshops were held, common guidelines were teased out, and a number of papers were commissioned. Altogether thirteen papers are being published by Saldru. Of these, nine are regional poverty profiles of the Eastern & Northern Transvaal, the PVW, Orange Free State & Qwa-Qwa, Kwazulu/Natal, Durban, Transkei, Ciskel, Port Elizabeth & Uitenhaga, and the Western Cape. The other four are cross-cutting studies focusing on Energy, Nutrition, Water Supply, and Housing.

One of the most important stages in the project was that of drafting the main questionnaire. Drawing largely on World Bank experience with similar surveys in other countries, a preliminary draft questionnaire {Mark One) was drawn up as a basis for discussion. A workshop in Cape Town involving some thirty social scientists and others from around South Africa took this draft apart and put it together again as Mark Two. This process of drawing upon a wide range of informed criticism and suggestions by means of commissioned comments and of workshops in different parts of country went on for almost a full year and took the questionnaire through at least twelve drafts, three of which were tested in pilot projects in the field. The final result was by no means perfect but the process did help

to eliminate a number of inconsistencies and to ensure that a lot of thought (and debate!) went into deciding what to include, what to exclude, and how best to phrase each question. Needless to say those working on the Project discovered more flaws after it was too late to change Mark Twelve but the lessons learned during the course of this first, base line, survey can be incorporated into subsequent surveys as the new South Africa develops an ongoing capacity to monitor living standards and the emerging pattern

of development.

Drawing up the integrated questionnaire was one part of the process. No less difficult was that of administering it, particularly in so diverse a field as
South Africa. The Project was fortunate to be able to enlist the services of a number of professional survey organizations, each with different strengths, to apply the questionnaire in the field. The organizations which undertook the actual field-work were the Bureau of Market Research (Pretoria), Data Research Africa (Durban), Human Sciences Research Council (Durban), Mark Data (Pretoria) and Social Surveys (Johannesburg). in addition a team in Umtata bad by Mr. Sintu Mpambani, in dose liaison with Saldru, worked through the Transkei.

In order to ensure consistency a number of workshops were held to train field workers in different parts of the country. Members of the Project staff based in Cape Town, kept in close touch with the main offices of the survey organizations in different centers. The months during which this took place were not the quietest in South Africa's history and we should like to pay a special tribute to those data gatherers in the field who were prepared to take considerable risks in order to do their work. The hijacking of one minibus containing a set of already completed questionnaires was a sharp reminder of the difficulties facing survey organizations. In the event only two of the 360 clusters chosen in the sample were not surveyed because of the dangers involved. A process was also put into place whereby observers independent of the particular survey organization working in an area were sent into the field to spot check the answers filled in for individual questionnaires. Tho process of verification in the field, whilst not as comprehensive as we should have liked, helped to confirm the accuracy of the household roster in most (though not quite all) areas of the country whilst at tho same time alerting us to certain weaknesses (particularly with regard to some of the anthropometric data).

Once collected, the information gathered had to be entered into a computerized data base and then cleaned. This process involving meticulous attention to detail took several months. It is this set of data, the first based on a comprehensive sample of the entire South African population (including the former TBVC--see footnote 1--states) and using an integrated
household questionnaire, that is now available to all who wish to make use of it. But access to data in South Africa is not sufficient unless it is accompanied by a deliberate process of ensuring that those who might find the data useful for whatever purpose have acquired the skills to analyze it. To this end, plans have been made to ensure that publication of the data is followed by a series of workshops in a training programme aimed at those in government, in trade-unions, in policy-making bodies, in universities, in nongovernmental institutions, and elsewhere. The purpose of these workshops is to ensure that, as far as possible, the data is disseminated in such a way that it becomes used to its highest potential as a basis for public policy debate in this country.

footnote 1: Transkei, Bophuthatswana, Venda and Ciskel. The other six non-independent homelands were also included in the sample.

METHODOLOGY
THE QUESTIONNAIRES

The main instrument used in the survey was a comprehensive household questionnaire. This questionnaire covered a wide range of topics but was not intended to provide exhaustive coverage of any single subject. In other words, it was an integrated questionnaire aimed at capturing different aspects of living standards. The topics covered included demography, household services, household expenditure, educational status and expenditure, remittances and marital maintenance, land access and use, employment and income, health status and expenditure and anthropometry (children under the age of six were weighed and their heights measured).

This questionnaire was available to households in two languages, namely English and Afrikaans. In addition, interviewers had in their possession a translation in the dominant African language/s of the region.

A crucial concept in the questionnaire was the definition of the household. The household definition was drawn up in such a manner as to avoid double-counting of individuals who may live in more than one place. Two definitions of the household were used. The first was used only in the first section of the questionnaire, i.e. the Household Roster and the second was used for the rest of the questionnaire. The first definition of the household comprised individuals who:

(I) live under this 'roof' or within the same compound/homestead/stand at least 15 days out of the past year, and

(ii) when they are together they share food from a common source (i.e. they cook and eat together); and

(iii) contribute to or share in, a common resource pool (i.e. they contribute to the household through wages and salaries or other cash and in-kind income or they may be benefitting from this income but not contributing to it, e.g. children, and other non-economically active people in the household. Visitors were excluded from this definition.

The second definition of the household included only those members who had lived "under this roof for more than 15 days of the last 30 days". This definition was derived to eliminate double-counting of individuals.

In addition to the detailed household questionnaire referred to above, a community questionnaire was administered in each cluster of the sample. The purpose of this questionnaire was to elicit information on the facilities available to the community in each cluster. Questions related primarily to the provision of education, health and recreational facilities. Furthermore there was a detailed section for the prices of a range of commodities from two retail sources in or near the cluster: a formal source such as a supermarket and a less formal one such as the "corner cafe" or a "spaza". The purpose of this latter section was to obtain a measure of regional price variation both by region and by retail source. These prices were obtained by the interviewer. For the questions relating to the provision of facilities, respondents were "prominent" members of the community such as school principals, priests and chiefs.

SAMPLING

The sample design adopted for the study was a two-stage self-weighting design in which the first stage units were Census Enumerator Subdistricts (ESDs, or their equivalent) and the second stage were households.

The advantage of using such a design is that it provides a representative sample that need not be based on accurate census population distribution. in the case of South Africa, the sample will automatically include many poor people, without the need to go beyond this and oversample the poor. Proportionate sampling as in such a self-weighting sample design offers the simplest possible data files for further analysis, as weights do not have to be added. However, in the end this advantage could not be retained and weights had to be added. (See below.)

The sampling frame was drawn up on the basis of small, clearly demarcated area units, each with a population estimate. The nature of the self-weighting procedure adopted ensured that this population estimate was not important for determining the final sample, however. For most of the country, census ESDs were used. Where some ESDs comprised relatively large populations as for instance in some black townships such as Soweto, aerial photographs were used to divide the areas into blocks of approximately equal population size. In other instances, particularly in some of the former homelands, the area units were not ESDs but villages or village groups.

In the sample design chosen, the area stage units (generally ESDs) were selected with probability proportional to size, based on the census population. Systematic sampling was used throughout that is, sampling at fixed interval in a list of ESDs, starting at a randomly selected starting point. Given that sampling was self-weighting, the impact of stratification was expected to be modest. The main objective was to ensure that the racial and geographic breakdown approximated the national population distribution. This was done by listing the area stage units (ESDs) by statistical region and then within the statistical region by urban or rural. Within these sub-statistical regions, the ESDs were then listed in order of percentage African. The sampling interval for the selection of the ESDs was obtained by dividing the 1991 census population of 38,120,853 by the 300 clusters to be selected. This yielded 105,800. Starting at a randomly selected point, every 105,800th person down the cluster list was selected. This ensured both geographic and racial diversity (ESDs were ordered by statistical sub-region and proportion of the population African). In three or four instances, the ESD chosen was judged inaccessible and replaced with a similar one.

In the second sampling stage the unit of analysis was the household. In each selected ESD a listing or enumeration of households was carried out by means of a field operation. From the households listed in an ESD a sample of households was selected by systematic sampling. Even though the ultimate enumeration unit was the household, in most cases "stands" were used as enumeration units. However, when a stand was chosen as the enumeration unit all households on that stand had to be interviewed.

Census population data, however, was available only for 1991. An assumption on population growth was thus made to obtain an approximation of the population size for 1993, the year of the survey. The sampling interval at the level of the household was determined in the

following way: Based on the decision to have a take of 125 individuals on average per cluster (i.e. assuming 5 members per household to give an average cluster size of 25 households), the interval of households to be selected was determined as the census population divided by 118.1, i.e. allowing for population growth since the census. It was subsequently discovered that population growth was slightly over-estimated but this had little effect on the findings of the survey.

Individuals in hospitals, old age homes, hotels and hostels of educational institutions were not included in the sample. Migrant labour hostels were included. In addition to those that turned up in the selected ESDs, a sample of three hostels was chosen from a national list provided by the Human Sciences Research Council and within each of these hostels a representative sample was drawn on a similar basis as described above for the households in ESDs.

DATA COLLECTION

Data collection was carried out by the survey organizations listed earlier. The workload and areas were assigned to the organizations on the basis of their previous experience and their geographical location. The Bureau of Market Research was responsible for the rural and the predominantly non-African urban areas of the Transvaal excluding the homelands. Mark Data conducted surveys in the Orange Free State, Qwa-Qwa, Bophuthatswana and Lebowa. Social Surveys covered the African townships in the PWW as well as Venda, Gazankulu and Kwandebele. Data Research Africa from Natal was responsible for the field work in Kwazulu and Kangwane. The rest of Natal and the Ciskei was covered by the HSRC in Durban. The HSRC in Cape Town covered the Northern, Western and Eastern Cape. Finally, a team under Sintu Mpambani from the University of the Transkei covered the difficult terrain in the Transkei.

Completed questionnaires were sent to Saldru where data entry management and cleaning were centralized.

DATA ENTRY, DATA MANAGEMENT AND CLEANING

All the questionnaires were checked when received. Where information was incomplete or appeared contradictory, the questionnaire was sent back to the relevant survey organization. As soon as the data was available, it was captured using local development platform ADE. This was completed in February 1994.

Following this, a series of exploratory programs were written to highlight inconsistencies and outlier. For example, all person level files were linked together to ensure that the same person code reported in different sections of the questionnaire corresponded to the same person. The error reports from these programs were compared to the questionnaires and the necessary alterations made. This was a lengthy process, as several files were checked more than once, and completed at the beginning of August 1994.

In some cases questionnaires would contain missing values, or comments that

the respondent did not know, or refused to answer a question. These responses are coded in the data files with the following values:

VALUE	MEANING
-1	The data was not available on the questionnaire or form
-2	The field is not applicable
-3	Respondent refused to answer
-4	Respondent did not know answer to question

WEIGHTS

A self-weighting sample design should in principle eliminate the need for weighting. A number of factors intervened, however, which made it essential to use weights after all. Amongst these was violence, which prevented survey teams from conducting interviews in two clusters on the East Rand; failure to continue interviewing in a cluster until the required take had been interviewed; and systematic under-representation of whites in the sample. This last problem resulted both from systematic non-response (whites were found to be more likely to refuse to be interviewed, or to be absent than other groups) and from sampling problems themselves.

The importance of race in determining living standards in South Africa is such that the racial distribution of the population has a major bearing on measures of living standards and inequality. It was thus regarded as essential that the problems mentioned above should be overcome by applying appropriate weights to the data. The most appropriate weights to apply would usually be the average values obtained in a cluster for the missing questionnaires from that cluster in order to capture the homogeneity usually inherent in residential contiguity. However, that presented some difficulty for the two clusters in which violence prevented surveying and for those clusters in which there were only a small number of questionnaires completed. It was felt that this method would therefore not be appropriate.

Accordingly it was decided to use weights as far as possible at the level of the old provincial/homeland boundaries and race. The listing of households in each cluster combined with the sampling interval was used to determine how many households should have been interviewed. Where this deviated from the number actually interviewed, this was taken into account. The assumption was that the households left out were racially distributed in the same proportion as the actual households interviewed. When these numbers were then calculated at the provincial level, a weight could be calculated for each race group to rectify errors made in the field work. These errors typically resulted from the fact that most field work organizations involved had little experience of using anything but a weighted sample and were used to replacements that could easily be added ex post, not necessarily in the same area. When these mistakes were discovered, it was too late to go back to the field.

The sample of 360 clusters of 25 households each based on an expected household size of 5 should have yielded a resident population of 45,000. In fact, a different household size should not affect the results. In any particular cluster, the expected take of individuals would remain the same if the census population were accurate, irrespective of household size, for a smaller household size (as in the case of whites) would only have yielded

more households, of whom a given proportion would have been interviewed. If in a particular cluster the census population was 472, every fourth household should have been interviewed (based on a sampling interval calculated to produce 125 persons per cluster in 1993, the expected take based on the census data of 118.1 per duster divided into the same population size). Irrespective of household size, then, one quarter of the cluster population would have been included in the survey. An average household size of 5 would have given 94 households of whom 23 would have been interviewed, i.e. 115 resident household members would have been found. If the household size were only three, on the other hand, one-quarter of the 157 households would have been 39, representing 117 household members. Only small differences from the expected take of 118 should thus arise, due to rounding. Only if the estimate of population based on the census is wrong, however, would the actual number of households deviate substantially from the expected take. In such a case, one quarter of the actual (i.e. listed or enumerated) rather than of the census population would have been included in the survey, i.e. there would have been an automatic adjustment. This gives the sample design its self-weighting character.

The census population for the survey data was estimated by applying Sadie's population growth rates to the adjusted 1991 census figures. The resultant racial and geographic distribution of the population of 40.1 million was presuming, of course, that no migration across provincial and homeland boundaries had occurred since the census. This implies that a raising factor of 891.4154 (40.1 million divided by an expected take of 45,000) should be applied to the results weighted by enumeration to obtain the population it represents. Applying the weights according to enumeration, 38.1 million people were covered by the survey, i.e. there was a 2 million under-enumeration amounting to about 5 per cent. Broken down by race, the under-enumeration was particularly large amongst whites, for whom the best census data exists, indicating that the problem did not lie so much with the census as with the survey. However, this is to be expected - a survey of this nature is better at capturing inequality and living standards than population size. Nevertheless, the margin of error in aggregate population estimates is relatively small, considering the presence of some homeless people, uncertainties about ESD boundaries in some areas and the likelihood of incomplete listings of households for various reasons. These results are therefore encouraging regarding the accuracy of the survey and also confirm that the adjusted census does not deviate substantially from population estimates obtained in a different manner.

However, the raised enumeration results deviate more from the census results where the provincial breakdown is concerned. The reason for this is not hard to find. The sample design introduced stratification only by geographic area (statistical regions) and proportion of the ESD population that was black. South African population clusters are still predominantly racially homogeneous, inter alia, because of past controls on residential patterns. It is therefore not surprising that in particular regions too few or too many clusters of a particular group were selected. In Natal, for instance, Coloureds and Indians are over represented in the data, even when weighted by enumeration, while Whites are under-represented. At the aggregate level, this should have little effect on the validity of the conclusions drawn, but it emphasizes the fact that care should be taken when drawing implications from the survey for Small populations. In small provinces (for instance, the new Northern Cape),

only a small number of clusters has been included, with the result that little can be concluded about living standards there, even though these clusters are important in determining overall distribution.

As a final comment on weights, the data provided for the user contains weights to correct for the enumeration difficulties discussed above as well as census-based
weights. If the user of the data wishes to use these weights they are found in the data file named "weight02". The variable name for the enumeration-based weight is "rsweight" and the name for the census-based weight is "rcweight". (Do not use the "sweight" and "cweight" variables.)