University of Michigan | University of Cape Town | Princeton University

THE CAPE AREA PANEL STUDY:

Overview and Technical Documentation Waves 1-2-3-4 (2002-2006)

David Lam, Cally Ardington, Nicola Branson, Anne Case, Alicia Menendez, Murray Leibbrandt, Jeremy Seekings and Meredith Sparks

October 2008

This is a working document, subject to revision

This draft working document (subject to revision) is co-authored by David Lam, Cally Ardington, Nicola Branson, Anne Case, Murray Leibbrandt, Alicia Menendez, Jeremy Seekings and Meredith Sparks.

Papers using the CAPS Waves 1-2-3-4 data should include the following acknowledgement:

The Cape Area Panel Study Waves 1-2-3 were collected between 2002 and 2005 by the University of Cape Town and the University of Michigan, with funding provided by the US National Institute for Child Health and Human Development and the Andrew W. Mellon Foundation. Wave 4 was collected in 2006 by the University of Cape Town, University of Michigan and Princeton University. Major funding for Wave 4 was provided by the National Institute on Aging through a grant to Princeton University, in addition to funding provided by NICHD through the University of Michigan.

We recommend the following citation for papers using CAPS Waves 1-2-3-4:

Lam, David, Cally Ardington, Nicola Branson, Anne Case, Murray Leibbrandt, Alicia Menendez, Jeremy Seekings and Meredith Sparks. *The Cape Area Panel Study: A Very Short Introduction to the Integrated Waves 1-2-3-4 Data*. The University of Cape Town, October 2008.

All documentation and survey instruments are available on the CAPS website: http://www.caps.uct.ac.za/

Preface

This document contains information on the design of the Cape Area Panel Study, sampling and structure of the questionnaires and datasets for Waves 1, 2, 3 & 4 (2002, 2003-2004, 2005, 2006).

The Cape Area Panel Study (CAPS) is a longitudinal study of the lives of youths and young adults in metropolitan Cape Town, South Africa. The first wave of the study collected interviews from about 4800 randomly selected young people age 14-22 in August-December, 2002. Wave 1 also collected information on all members of these young people's households, as well as a random sample of households that did not have members age 14-22. A third of the youth sample was re-interviewed in 2003 (Wave 2a) and the remaining two-thirds were re-visited in 2004 (Wave 2b). The full youth sample was then re-interviewed in both 2005 (Wave 3) and 2006 (Wave 4). Wave 3 also includes interviews with approximately 2000 co-resident parents of young adults. Wave 4 also includes interviews with a sample of older adults (all individuals from the original 2002 households who were born on or before 1 January 1956) and all children born to the female young adults. The study covers a wide range of outcomes, including schooling, employment, health, family formation, and intergenerational support systems.

CAPS began in 2002 as a collaborative project of the Population Studies Center in the Institute for Social Research at the University of Michigan and the Centre for Social Science Research at the University of Cape Town (UCT). Other units involved in subsequent waves include UCT's Southern African Labour and Development Research Unit and the Research Program in Development Studies at Princeton University. Primary funding is provided by the National Institute of Child Health and Human Development of the U.S. National Institutes of Health (NIH). Additional funding has been provided by the Office of AIDS Research, the Fogarty International Center, and the National Institute of Aging of NIH, and by grants from the Andrew W. Mellon Foundation to the University of Michigan and the University of Cape Town.

Documentation for CAPS Waves 1-2-3-4:

The Cape Area Panel Study: Overview and Technical Documentation for Waves 1-2-3-4 Introduces the major motivations for the CAPS project as a whole, the project team and sponsors, details of the original sample design, as well as describing fieldwork, training, the survey instruments, and response rates for each of Waves 1, 2, 3 and 4.

A Very Short Introduction to the Integrated Waves 1-2-3-4 (2002-2006) Data: This document is intended to familiarize analysts with the organization of the public release data sets.

CAPS Waves 1-2-3-4 Panel variables crosswalk: Matches variables which are repeated across the panel in merged datasets.

Contents

1. Introduction

The Cape Area Panel Study (CAPS) follows the lives of a large and representative sample of adolescents in Cape Town as they undergo the multiple transitions from adolescence to adulthood. The study commenced in 2002, with approximately 5250 households and 4750 young people between the ages of 14 and 22 interviewed as part of Wave 1 of CAPS. In 2003 and 2004, these young adults were re-interviewed in Waves 2a and 2b of the project. In Wave 3 (2005) we re-interviewed the entire young adult sample along with a questionnaire for their households. In Wave 4 (2006) we have re-interviewed the entire young adult sample and their households for a fourth time, also adding a sample of adults aged 50 and over and a short questionnaire covering all children of female young adults.

Together, this series of interviews will constitute a significant source for the study of adolescents in post-apartheid South Africa. CAPS covers a range of aspects of adolescence, including especially schooling, entry into the labour market (i.e. employment, unemployment and job search), sexual and reproductive health, and familial support. In addition to the data collected from the young people themselves, parents and other older household members, and we can combine the CAPS data on individuals and households with community- and school-level data.

But CAPS is not simply a study of adolescents or adolescence. Because the patterns of inequality in society as a whole are rooted in the differentiation evident or generated in this age span, CAPS is a study also of transition - and the lack of change - in the 'new' South African society as a whole. To what extent have the opportunities facing South Africans changed since the end of apartheid? What factors shape or determine whether South Africans end up rich or poor, healthy or sick, happy or unhappy? A growing number of studies in South Africa are concerned with the persistence of poverty over short periods of time. CAPS is concerned with how poverty and inequality are reproduced across generations.

This document provides detailed information on the CAPS sample design, response rates for Waves 1, 2, 3 and 4, weights to adjust for the sample design and non-response and the structure of the Waves 1, 2, 3 and 4 questionnaires.

For a very brief introduction and to familiarize yourself with the organization of the public release data sets, please see the *A Very Short Introduction to the Integrated Waves 1-2-3-4 (2006) Data.*

For a mapping of variables across waves, please see the *CAPS Waves 1-2-3-4 Panel Variable Crosswalk.*

1.1. Motivation and Goals of the CAPS Project

The CAPS project was designed to provide rich detail on the transitions made by young South Africans as they move through school, enter the labour force, begin sexual activity, move into their own households, and start their own families. Since most existing sources of data in South Africa only provided cross-sectional information on the lives of young people, one of the major objectives of the project was to launch a longitudinal survey that would follow the same respondents over time. Another important objective was to include detailed

information on the household environment and family connections of young people, including information on all other individuals living in the respondent's household.

CAPS was conceived in the period immediately following the end of apartheid in South Africa. The abolition of racial discrimination in education, employment, and health and welfare policies, and of racial segregation in where people could live, promised to open up new opportunities and incentives for the post-apartheid generation of young people. But this generation was growing up in a society beset by other, worsening challenges: unemployment, crime and violence, and the HIV/AIDS pandemic. How would the home and family environment, the neighbourhood, the school and the labour market combine to shape the routes that young people followed and their eventual destinations?

The relations between home and neighbourhood, schooling and employment outcomes are the subject of study in many societies other than South Africa. It was anticipated that the South African case would be of broader relevance. Whilst South Africa's political changes may have been distinctive, the parallel processes of social and economic change were common to many developing and transitional countries. Secondly, the particular circumstances of changing public policy meant that South Africa serves at the same time as an unusual laboratory. For example, the end of segregation in the schooling system resulted in a possibly unique shift in the range of choices that parents and students from a range of social and economic backgrounds could make about schooling. Thirdly, the study of adolescent decision-making in South Africa held out the possibility of valuable lessons for even the more developed countries of the world. For example, the effects of extreme income inequality on the expectations and behaviour of young people in South Africa might provide lessons for countries such as the USA, which have themselves experienced growing inequality.

The conceptualization and design of CAPS took place at a time of rapidly improving availability of data on South Africa. Under apartheid there was little or no systematic collection of data on the circumstances, attitudes or behaviour of the black majority of the South African population. It was only in 1993 that the country's first comprehensive, countrywide survey was conducted of household income and expenditure; this was the Project for Statistics on Living Standards and Development (PSLSD), run by the University of Cape Town together with the World Bank. From 1993 there was an explosion of data, much coming from the reformed parastatal statistics agency Statistics South Africa, and some coming from university-based research initiatives (Seekings, 2001).

The new cross-sectional data-sets – such as the 1993 PSLSD and the subsequent October Household Surveys (OHSs) conducted by Statistics SA – generated a range of findings on the challenges faced by young people in South Africa, and on the consequences of these in later life. Anderson, Case and Lam (2001) summarized the findings on education from cross-sectional household surveys such as the PSLSD and OHSs. They showed that racial differences in educational attainment had steadily declined under apartheid, although even in the early 1990s the proportion of white adolescents passing the school-leaving or matriculation examination was more than double the corresponding proportion of African adolescents. In 1995, white adolescents were on average two grades ahead on their African counterparts by the age of seventeen. Unusually for the developing world, there was no corresponding gender gap in schooling. Anderson *et al.* show that the racial schooling gap did not result primarily from lower enrollment rates or higher dropout rates, and certainly not because African adolescents were leaving school to find work. Rather, they suggest, it was

due to grade repetition, i.e. African adolescents were failing and repeating grades in large numbers. Unfortunately, the cross-sectional data-sets assembled in the 1990s did not contain the detailed longitudinal data on progress through school that would allow for careful analysis of the causes and consequences of grade repetition. Anderson *et al*. also show that there is a correlation between an adolescent's mother's education and the adolescent's own progress through school. Family background seems to matter. But, as they point out, 'it is not clear what causal mechanisms drive this relationship' (48). One of the limits of surveys such as the OHSs was that they only collected data on co-resident kin (parents, grandparents, even children); this made it especially difficult to assess the dynamic effects of family and home background.

Cross-sectional data-sets revealed also the clear and close correlation between education and, in later life, earnings, but they could shed little light on the process through which adolescents entered the labour market. Why did some adolescents drop out of school without passing matric, how did they search for employment, and what kinds of jobs did they get? What kinds of factors made it more likely that an adolescent, with any given education, would secure employment, and especially secure and well-paid employment? Cross-sectional data showed high rates of labour market participation among young women, including young mothers, but not precisely how having children affected job search and employment, or what factors made it easier for young mothers to combine employment with motherhood.

In a review of the new quantitative data on adolescence, Bray (2002) contrasts how much is known about adults – what they do, what they earn, what they think, how good is their health, and so on – with how little is known about children and especially adolescents. Very little is know about schooling itself, nothing about how adolescents enter into the labour market, and very little about their health, sexual behaviour or experiences of pregnancy and parenthood. When something is known about children or adolescents, the information is generally collected from adults; the voices of young people themselves are rarely heard.

New research on adolescence in South Africa clearly needed to transcend the existing limits in a number of ways. First, data needed to be collected on a range of topics – including schooling, employment, health and relationships – that had been neglected as far as young people are concerned. Secondly, young people need to be understood within a wider range of relationships than simply the co-residential household. Data is required on relationships with other kin, including especially those close kin who were not co-resident with the adolescent. Finally, data needed to be collected on a longitudinal basis, at least through more thorough retrospective questions and ideally through a panel study. The Cape Area Panel Study was conceived and designed to achieve these objectives.

The age range of 14 to 22 was chosen as the target age range for the initial cohort to be recruited into the study. This was viewed as narrow enough to ensure that reasonably large samples would be available at each age, but broad enough to cover a broad range of transitions. Issues of school enrolment and school progress would be most important to those in the young end of the sample, while issues of work, reproductive health, and family formation would be more important at the older end. Although the CAPS project is focused primarily on young adults, there was an interest in collecting data on individuals at other ages as well. Data on other members in the households where young adults were living would be important for understanding the household and family environment affecting the lives of young people. Data on households without young adults would be valuable for

understanding the connections of adults to their non-resident children and would greatly increase the value of the data by providing a representative sample of the Cape Town population.

CAPS was not the first panel study to be conducted in South Africa. Notable predecessors include the Birth-to-Ten birth cohort study in the Johannesburg area, the KwaZulu-Natal Income Dynamics Study (KIDS) and the Durban-based Transition to Adulthood study.[1] The Birth-to-Ten (later extended into Birth-to-Twenty) study tracked the health and development of a cohort of children born in public clinics in the Johannesburg area between April and June 1990. The emphasis of the study was medical and psychological, given the infancy of the children through the early years of the study (see Barbarin and Richter, 2001). KIDS, in contrast, was explicitly focused on the economic and to some extent social aspects of households. The KIDS project revisited in 1998 and 2004 the households interviewed in 1993 for the KwaZulu-Natal part of the PSLSD. The PSLSD/KIDS panel thus comprised a three-wave panel of about 1,100 households (May and Roberts, 2001). A third notable panel study, Transitions to Adulthood, was much closer in design to CAPS. In late 1999, a sample of about 3,000 young adults between the ages of 14 and 22 were interviewed, three-quarters in the Durban metropolitan area and one-quarter in a more rural district in northern KwaZulu-Natal. The study was concerned primarily with exposure to sex education, knowledge of HIV/AIDS and sexually-transmitted diseases (STDs), sexual and reproductive histories, and contraceptive use (Rutenberg *et al.*, 2001). Two years later, in late 2002, most of the respondents were re-interviewed. The Transitions study is thus a two-wave panel. Each of these three panel studies has generated invaluable data and spawned important analyses of South Africa's changing society. These panels provided useful lessons, informing the design and subsequent implementation of CAPS, which was designed to build on the experiences of these previous studies while moving forward to cover new ground.

1.2. Overview of Waves 1, 2, 3 and 4

The first interviews for CAPS were conducted in early August, 2002. The fieldwork in this wave 1 comprised the administration of 'young adult' and 'household' questionnaires as well as the evaluation of the literacy and numeracy of the young adults in the sample.

Wave 2 was conducted over the period July 2003-December 2004, split into two separate fieldwork operations, Wave 2a (2003) and Wave 2b (2004). The goal of Wave 2 was first to keep in contact with CAPS Young Adult respondents who had completed Wave 1 questionnaires during the 3-year period before Wave 3, planned for 2005. Additionally, Wave 2 provided the opportunity to update areas of core panel data, such as employment, schooling and household rosters, as well as to introduce new modules to probe selected topics in further detail. This section describes fieldwork operations and the tracking of young adult respondents.

[1] Other longitudinal studies include the demographic surveillance studies in Hlabisa (northern KwaZulu-Natal) and Agincourt (in the Mpumalanga lowveld), as well as localised studies such as the HIV/AIDS panel run by Dr Frikkie Booysen of the University of the Free State, in Welkom and QwaQwa in the Free State. The Labour Force Survey (run twice per year by Statistics South Africa) is a rotating panel by design, but is difficult to analyse as a panel.

Wave 2 took place in two components in 2003 and 2004. The original plan for CAPS was to do two waves, one in 2002 and a follow-up in 2005. Additional funding provided by the NIH Office of AIDS Research and the Andrew W. Mellon Foundation made it possible to add partial waves with particular thematic focuses in 2003 and 2004. Approximately one-third of the young adult sample was re-interviewed in 2003 (Wave 2A), with the remaining two-thirds interviewed in 2004 (Wave 2B). Re-interviews in waves 2A and 2B provided an opportunity to update the schooling and employment data that was collected in 2002, as well as probe selected topics in additional detail and to add new topics not covered in Wave 1. Waves 2A and 2B included the following innovative components:

- A module on HIV/AIDS stigma (in wave 2A)
- Modules on employment and unemployment (in wave 2B)
- A module on school choice (in wave 2B)
- A new format for recording data on schooling and work, including a month-by-month calendar and job schedule

Wave 2A, in 2003, was focused primarily on the topics of sex and AIDS, including AIDS stigma. The questionnaire included questions that updated data collected in wave 1 together with new modules focusing primarily on attitudes toward HIV/AIDS.

Wave 2B, in 2004, also included a mix of repeat questions to update data from 2002 and new questions, focusing primarily on employment, unemployment and school choice (i.e. how and why students choose which school to attend). The new questionnaire included an expanded month-by-month calendar, on which was recorded on a monthly basis when the household was affected by shocks (such as serious illness, death, job loss, or accessing a new grant), when the respondent was studying, when he or she was working or looking for work, and so on.

Wave 3 of CAPS was conducted between April and December 2005. The target sample for Wave 3 was the full set of 4,750 young adults originally interviewed in Wave 1. Wave 3 also included a household questionnaire similar in design and execution to the household questionnaire used in Wave 1. In addition, a parent questionnaire was administered to a parent or guardian of each young adult whenever possible.

The primary focus of the Wave 3 young adult questionnaire was to update data on schooling, employment, pregnancies and births, and personal health. New components in Wave 3 included a detailed residential and schooling history, questions focusing on intergenerational transfers, time allocation, relationships with parents or guardians and a detailed history of all sexual partners.

The household questionnaire in Wave 3 was expanded from the Wave 1 content to include questions relating to family support, in particular the claims, obligations and responsibilities spanning large distances. Particularly relevant to CAPS are the links between migrant workers from the Eastern Cape Province supporting kin in rural areas through remittances. Additionally, the questionnaire includes expanded modules to uncover methods used by individuals and households to respond to the negative ‘shocks’ of poor health, death, or unemployment, as well as expected obligations in the event of positive ‘shocks’ such as getting a job, a better paying job or a new grant.

Wave 4 of CAPS was conducted between April and December 2006, with tracking conducted in early 2007. The target samples for Wave 4 are the following:

- The full sample of young adults (these are mostly aged 18-26 in 2006)
- The biological children of all female CAPS young adults
- All residents of original Wave 1 CAPS households who are age 50 or over in 2006

A household questionnaire was administered in each household that contained any of the above respondents.

In addition to providing follow-up information on the school, work, and childbearing histories of CAPS young adults, Wave 4 expands the focus on health and systems of family support.

1.3. Sponsors and Project Team

CAPS began in 2002 as a collaborative project of the Population Studies Center in the Institute for Social Research at the University of Michigan and the Centre for Social Science Research at the University of Cape Town (UCT). Other units involved in subsequent waves include UCT's Southern African Labour and Development Research Unit and the Research Program in Development Studies at Princeton University. The Principal Investigator for Waves 1-2-3-4 was David Lam, Professor of Economics and Research Professor in the Population Studies Center at UM. The co-principal investigator for Waves 1 and 2 was Jeremy Seekings, Professor of Sociology and Political Science at UCT and Director of the Social Surveys Unit in CSSR. The co-principal investigators for Wave 3 were Jeremy Seekings and Murray Leibbrandt, Professor in the School of Economics at UCT and Director of the Southern Africa Labour Development Research Unit. The co-principal investigators for Wave 4 were Murray Leibbrandt and Anne Case, Professor of Economics and Public Affairs at the Woodrow Wilson School of Public and International Affairs and the Economics Department at Princeton University.

Major funding for Wave 1-2-3 of CAPS was provided by Research Grants R01-HD-039788, "Families, Communities, and Youth Outcomes in South Africa," and R01-HD-045581, "Family Support and Rapid Social Change in South Africa," from the National Institute of Child Health and Human Development, part of the U.S. National Institutes of Health (Principal Investigator: David Lam). Additional funding was provided by grants from the Andrew W. Mellon Foundation to both the CSSR and PSC, supplemental funding from the Office of AIDS Research of the U.S. National Institutes of Health, and Research Grant D43-TW-000657, "Population Research and Training in Developing Countries," from the John E. Fogarty International Center of the U.S. National Institutes of Health (Principal Investigator: David Lam). Major funding for Wave 4 was provided by the National Institute on Aging through a grant to Princeton University (Principal Investigator: Anne Case), in addition to funding provided by NICHD through the University of Michigan.

CAPS involves a large and growing team of economists, sociologists, social anthropologists, demographers, educationalists and social psychologists. Waves 1, 2, 3 and 4 of CAPS were developed with input from researchers at UCT, UM, Princeton University and other institutions. These include Kermyt Anderson, Cally Ardington, Ann Beutel, Ann Biddlecom, Justine Burns, Anne Case, Owen Crankshaw, Malcolm Keswell, Murray Leibbrandt, Brendan

Maughan-Brown, Alicia Menendez, Cecil Mlatsheni, Nicoli Nattrass, Jolene Skordis, Volker Schoer, Joanne Stein, Matthew Welch, Francis Wilson and Martin Wittenberg. Valuable guidance on sample design was provided by Jim Lepkowski of the Survey Research Center at UM. Nick Taylor and Penny Vinjevold played a major role in development of the Wave 1 Literacy and Numeracy Evaluation.

Academic oversight and management of the fieldwork operations was provided principally by Jeremy Seekings for Waves 1, 2 and 3, David Lam for Waves 3 and 4, Murray Leibbrandt for Waves 3 and 4 and Anne Case for Wave 4.

CAPS staff who played a major role in data management and data processing include Christine Schippers, Meredith Sparks, Nicola Branson and Miguel Lacerda.

1.4. Consent and Confidentiality

The CAPS project operates under the approval of human subjects review boards at both UM and UCT. Ethical approval for Wave 4 was also granted by the human subject review board at Princeton University. Project staff and field interview teams receive training in issues of informed consent and confidentiality. Written consent is obtained from all respondents, and written parental consent is obtained for interviews with respondents under the age of 18.

Issues of confidentiality are given careful attention in preparing public release data sets. In addition to removing all names, addresses, and contact details, we also remove names of schools, names of employers, day of birth, and any other information that might compromise confidentiality. We also do not release the actual census enumeration area identification numbers, since these numbers can be linked back to specific neighbourhoods, some of which are quite small. We replace the original EA numbers with alternative EA numbers (*cluster*) that allow researchers to know which households come from a common EA, but which do not correspond to identifiable geographic areas. Some researchers have legitimate reasons to be interested in the specific schools attended by students, or the specific neighbourhoods in which they live. We will consider releasing restricted data sets containing selected restricted variables on a case-by-case basis, with additional security protections expected from the researcher.

1.5. Other Information about CAPS

Additional information about CAPS can be found in the following places.

- CAPS web site: http://www.caps.uct.ac.za: The CAPS web site contains all of the documents listed, as well as additional information about the CAPS project.
- *A Very Short Introduction to the CAPS Integrated Waves 1-2-3-4 Data*: This document provides information on the organization of the public release datasets including variable names and the integration of data from multiple waves.
- *CAPS Waves 1-2-3-4 Panel Variable Crosswalk*: This document provides a mapping of panel variables in the public release datasets.

2. Wave 1 Sample Design

Motivated by the broad project objectives described in section 1.1 above, the CAPS sampling plan was designed with a number of specific goals in mind:

1. The design should produce a household sample that when appropriately weighted would be a representative sample of households in metropolitan Cape Town at the time of the survey, including both households with and without young adult residents.
2. The design should produce a young adult sample that when appropriately weighted would be a representative sample of the non-institutionalized population aged 14-22 in metropolitan Cape Town at the time of the survey.
3. The design should produce a young adult sample in the range of 4,500-5,000 young adult respondents, and should include a large enough number of households without young adults to draw statistically meaningful inferences about those households.
4. The design should produce large enough samples of young adults from each of the three major population groups to make statistically meaningful statements about each separate group. This goal was operationalized by having a target young adult sample with roughly equal numbers of African and coloured young adults and a white sample roughly half as large.
5. The design should take maximum advantage of census and geocode information available at the time of the survey, balancing statistical precision with the pragmatic realities of fieldwork operations.
6. The sample design should include multiple young adults per household, with an upper limit that would avoid excessive burden on the household and on field resources.

Given these goals, a stratified two-stage sample was designed by working backwards from the target number of young adults in each of the three population groups. The first stage was the selection of sample clusters. The second stage was the selection of households within each cluster. Since the 2001 census was not yet available, the 1996 census was used as the basis for the sample design. The Enumeration Areas (EAs) from the 1996 census were used as the basic sampling unit for the first stage selection of clusters.

Metropolitan Cape Town, as defined in the 1996 Population Census, included 4,759 populated EAs. The population of metropolitan Cape Town in the 1996 census, using the weights provided by Statistics South Africa, was 2,554,674. This was 6.3% of the population of South Africa. The metropolitan Cape Town population was 26% African, 50% coloured, 1.5% Indian, and 22% white. The total South African population was 77% African, 9% coloured, 2.6% Indian and 11% white. The Cape Town population in the 2001 census was 32% African, 48% coloured, 1.5% Indian, and 19% white. Since the coloured population in the 1996 census was roughly twice the size of the African and white populations in Cape Town, the goal of equal sample sizes from the African and coloured group meant that the sample would be designed to select African households with roughly twice the probability of coloured households. Because white households were much less likely to contain young adults, white households also had to be substantially oversampled, even to produce a sample of young adults that was half the size of the African and coloured samples.

In drawing our sample we had access to both the 10% micro-sample of the 1996 Census and the computerized tabulations from the 100% sample that included aggregate statistics at the EA level. The 10% sample did not include geographic identifiers at the EA level, but was useful for simulating samples at the individual and household level for metropolitan Cape

Town. The tabulated data was useful for generating the number of households in each population group in each EA, an essential piece of information for our sample design.

Our target of oversampling African and white households was done by stratifying the sample based on the predominant population group living in each EA. The tabulated data from the 1996 census for each EA was used to calculate the percentage of household heads in each EA that were classified as African, coloured, and white. Each EA was then categorized by the predominant population group, where a simple plurality was sufficient to produce a given characterization. A sample of EAs was drawn separately from each of these three sets of EAs, with the goal of producing a target number of 14-22 year-olds from each population group.

Since it was impossible to know in advance which households would contain residents in the 14-22 age group, it was necessary to design a sample of "screener" households that could be expected to produce the target number of young adults. The 1996 census 10% sample for metropolitan Cape Town was used to project how many randomly selected households would produce a given number of 14-22 year-olds. Because the average number of 14-22 year-olds per household varied considerably by population group, the sampling design was adjusted accordingly.

2.1 Selection of Clusters

The first stage of selection was to select the sample clusters or Primary Sampling Units (PSUs), using the 1996 Census EAs as the basic building blocks. The method of selection used was that of Probability Proportional to Size (PPS), with the measure of size being the number of households in each EA as measured by the 1996 Population Census. This method was chosen as it provides the most efficient way to obtain equal sub-sample sizes across two stages of selection.

The EA sample was stratified by the EA population group composition. As discussed above, each EA was classified as African, coloured, or white based on a plurality of the household heads' population group. In choosing how many households to sample within each EA we took several factors into account. Larger numbers of households per EA meant lower cost of fieldwork, but implied a loss in statistical precision of the sample. Balanced against the loss in statistical precision at the population level was the fact that larger numbers of households per EA created potential advantages in the analysis of the impact of neighbourhood and school characteristics on various outcomes. We decided that 25 households per EA was an appropriate compromise among these considerations. In the case of EAs with fewer than 25 households, these small EAs were linked with larger neighbouring EAs to produce a primary sampling units (PSUs) with at least 25 households. The linking of EAs to produce PSUs with at least 25 households was done prior to the selection of PSUs for the sample.

Another choice that had to be made was how many households to interview that did not contain residents age 14-22. Although we wanted to include some of these households in the sample, it was clear from simulations with the census that a simple random sample of households would produce a higher proportion of households without young adults than would be cost effective given the goals of the project. We decided to aim for a target of selecting roughly 50% of the households without young adults in African and coloured areas, and roughly 30% of the households without young adults in white areas. The reason for

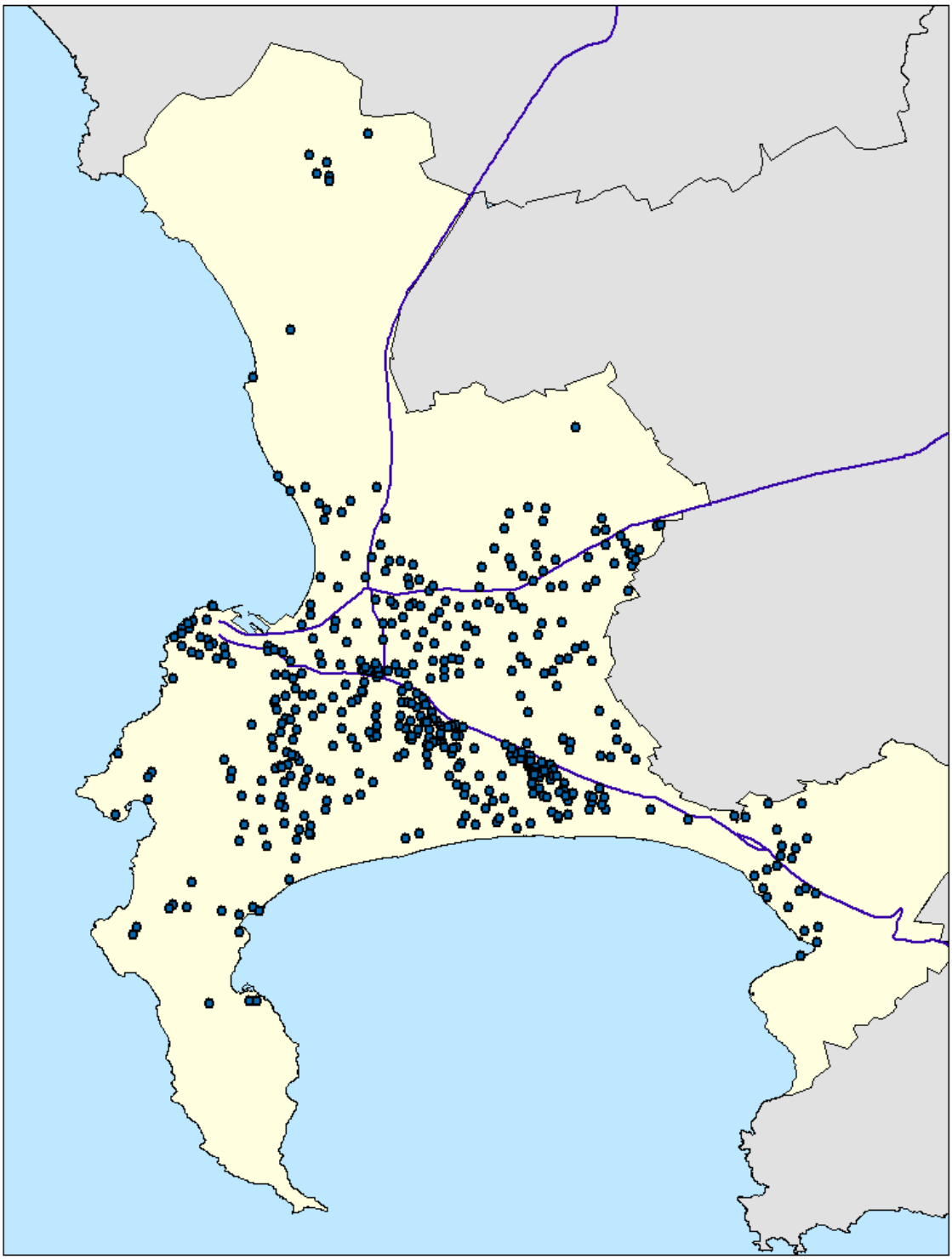
using a lower fraction in white areas was that a significantly lower fraction of households would be expected to contain young adults in white areas (roughly 50% of households in African and coloured areas were estimated to have residents age 14-22, compared to only about 25% of white households). Choosing half of the white households without young adults would have produced more households without young adults than we needed, at the cost of diverting resources away from interviewing young adults.

Taking all of these factors into account, we generated a target number of screener households in each stratum. These targets were roughly 3,200 African, 3,000 coloured, and 4,300 white screener households. Given a) the target number of households in each stratum; b) the target of 25 households per PSU; c) the actual number of households in each stratum; and d) the number of PSUs in each stratum, we drew a random sample of PSUs in each stratum. PSUs were chosen with probably proportional to size, meaning that if PSU *X* had twice as many households as PSU *Y*, the probably of selecting *X* was twice as high as the probability of selecting *Y*. The final number of EAs in the CAPS sample was 440, about 10% of the EAs in Cape Town in the 1996 census. Figure 1 shows the location of our enumeration areas (the locations are approximate in order to protect confidentiality of respondents). As can be seen in the map, our enumeration areas cover the full geographical range of metropolitan Cape Town.

2.2 Selection of Households

In the second stage of the sample selection, households were selected in each of the PSUs selected for the sample. This was done using aerial photographs (orthophotos) of each EA, where, as noted above, more than one EA may have been included in a given PSU. CAPS project staff took each aerial photograph, attached a transparent cover sheet, outlined every dwelling in the EA on the cover sheet, and then numbered each dwelling. The basic procedures followed the methods described in Crankshaw et al. (2001). Because the selection of EAs was based on 1996 census data – data that was almost six years old when the sample was being drawn – updating of information from EAs was essential. The aerial photographs were more current, and thus provided one important source of updating. In addition, field teams were sent out to do on-site inspections of most EAs, including all EAs that appeared to be in transitional areas or that had features that could not be clearly identified. These field teams updated the EA listings that were made from the aerial photographs, providing information such as identification of dwellings that were not residences, providing specific details regarding apartment buildings, and noting areas in which houses had been destroyed or new houses had been built. Most of the information provided by these field teams was based on walking or driving through the EA, without knocking on doors.

Figure 1. CAPS Enumeration Areas



Once the listing for a PSU had been updated, 25 households were selected at random from the numbered residential units on the aerial photograph (or combined aerial photographs in cases in which more than a PSU consisted of more than one EA). This was done by choosing a random start point and a skip interval based on the number of dwellings in the PSU (the skip interval was simply the number of dwellings in the PSU divided by 25). The selected 25 households were marked on the transparent cover sheet. The aerial photographs, cover sheets, notes from on-site updating, and detailed street maps were provided to the field interview teams.

Secondary households: While our procedures for selecting households should have been effective in producing a random sample of 25 households from each selected PSU, we were concerned that "backyard shacks" and other types of secondary household units might be missed. These secondary units might not be apparent in either the aerial photographs or in on-site visits to the EAs. Since backyard shacks are common in many township areas, and since we did want them to be underrepresented in the final sample, we took additional steps to make sure that backyard shacks were included in the sample. On the screening form administered to selected households, interviewers were advised to ask respondents the following question: "Are there any other separate residences in this property – for example, a backyard dwelling, a separate servant's quarters, or a separate flatlet, or is there more than one household living under this roof?" A household was considered a separate household if they "do not eat together or out of a common pot." In these cases the separate household was added to the sample and given its own household identification number. Interviewers were then advised to administer the standard screening questionnaire to the separate household.

Our procedure for adding secondary household units to the sample should have helped solve the problem of under-representing such units in the sample. There was some risk that it could have produced an oversample of such units, however, since some of the units identified as secondary units could have been previously identified as a separate residence on the aerial photograph. This would mean that these dwellings had two opportunities to be selected, giving them twice the probability of being selected as primary dwelling units. While there was no checking for this possibility in the field, we were able to check this *ex-post* by comparing the reports of secondary units with the original field maps. Preliminary analysis suggests that this was not a serious problem, but further investigation is underway.

2.3 Screening and Household Selection in the Field

When interview teams were sent into selected PSUs, the first step was to locate the selected screener households using the aerial photographs, street maps, and field notes. In most areas an advance letter from the project director at UCT was distributed to the selected households to inform them of the purpose of the interview and request cooperation with field staff. A screening form was prepared for each of the 25 selected screener households. These forms included the household number assigned by CAPS staff before the sample was selected, a screener number from 1 to 25, and questions about the total number of household members and the number of members aged 14-22. When interviewers made contact with an adult household member they informed them of the purpose of the survey and asked them to provide the information on the screener form.

If households had resident members aged 14-22, the household was automatically considered part of the CAPS sample. The interviewers then proceeded to the household questionnaire

and asked to speak to the adult household member who was most knowledgeable about everyone in the household. This person was informed about the purposes of the study and was asked to sign a written consent form. The household questionnaire was completed, followed by administration of the young adult questionnaire to up to three household members age 14-22.

If households had no resident members aged 14-22, the screener form directed the interviewers to follow one of two rules. In areas identified as African and coloured (based on the predominant population group in the EA, not the actual characteristics of the particular household), households were to be selected as part of the CAPS sample if the screener number was even (that is, ended in 0, 2, 4, 6, or 8). In these cases the interviewers would proceed to the household questionnaire. If the screener number was odd the household was not selected into the sample. In white areas the rule was that a household was selected into the sample if the screener number ended in 3, 6, or 9. As noted above, a smaller fraction of households without young adults was selected in white areas because of the lower prevalence of households with young adults in those areas. Since the screener numbers went from 1 to 25, these rules imply that among households with no residents aged 14-22, 48% of African and coloured households and 28% of white households would be selected in the sample.

Independent of the screener number or the number of young adults, interviewers asked whether there were any other household units on the property. If it was reported that there was an additional household on the property, a new household screener form was generated for this household. The new household was given the same household number as the original household, but an additional digit was added at the beginning. For example, if the original household was number 124, the first secondary household on that property was numbered 1124, the second was numbered 2124, etc. The new households were given the same screener number as the original household, and the screening procedures were followed in the same way as they were for the primary household. Thus, a secondary household with a resident aged 14-22 was automatically added to the CAPS sample, and a secondary household without a resident aged 14-22 was added only if the screener number satisfied the appropriate number rule for that area. About 12% of the households in our final screener sample were secondary households.

The addition of secondary households into the sample could increase the number of sampled households in an EA above the 25 called for in the sample design. An EA could potentially have had 50 or more households selected into the sample. On the assumption that the 1996 census included the secondary households in its count of households, while our listing based on aerial photographs did not, this implies that households in this EA had twice the probability of selection as households in an EA with only 25 selected households. This is taken account of in construction of the sample weights.

2.4 Selection of young adults

The sample design called for young adults age 14-22 to be selected into the sample based on the household roster that would be collected during the household interview (see the discussion of questionnaires below). We considered it highly desirable to include multiple young adults per household whenever possible. This would make it possible to look at issues such as gender differences within households, effects of birth order, and differences in outcomes between biological children, step-children, other relatives, and non-relatives. We

wanted to put a limit on the number of young adults that would be selected in any given household, however, for a number of reasons. Interviewing large numbers of young adults in a single household would put a large response burden on the household, would tie up field resources that could be used in other households, and would complicate the design of survey instruments, with relatively limited research payoff. Analysis of census data suggested that relatively few households would have more than three residents age 14-22. Based on all these considerations, we decided to set an upper limit of three young adults per household. In cases in which there were more than three young adults per household, the three with the most recent birthdays would be selected.

Summary: Here is a summary of the key points of our sample design:

- The sample was stratified on the predominant population group of the census enumeration area, with strata for the three major population groups in Cape Town – African, coloured, and white.
- EAs with fewer than 25 households were combined with nearby EAs to produce primary sampling units with at least 25 households.
- A sample of PSUs was selected within each stratum with probability proportional to size. The probability of selection was roughly twice as high in African and white areas as in coloured areas. This was based on a target of producing roughly equal numbers of African and coloured young adults, and about half as many white young adult respondents.
- Within each PSU a sample of 25 screener households was drawn using aerial photographs combined with on-site inspection and updating.
- Secondary households such as backyard shacks on the same property as screened households were added to the screened sample and treated in the same way as all other screened households.
- All screened households with members aged 14-22 were selected into the final sample of interviewed households.
- Households without any members aged 14-22 were selected into the final sample with probability around 0.5 in African and coloured areas and with probability around 0.3 in white areas.
- Up to three young adults were selected for the young adult sample from each household. In cases where there were more than three young adults, the three with the most recent birthdays were selected.

3. Sample selection for Waves 2, 3 and 4

3.1. Wave 2a and 2b Young Adult samples

In Wave 2 the aim was to re-visit roughly one third of the sample in 2003 (Wave 2a) and the remaining two-thirds in 2004 (Wave 2b). The Wave 2a sample was selected as follows.

- For enumerator areas classified as predominately African, every second enumerator area between Gugulethu and Macassar was selected.
- For enumerator areas classified as predominately Coloured, every third enumerator area was selected.
- Every third white young adult was selected regardless of the enumerator area.

3.2. Wave 3 and 4 Young Adult samples

The target sample for Waves 3 and 4 was the full set of 4,752 young adults originally interviewed in Wave 1, except those known to be deceased or mentally ill from fieldwork in previous waves.

3.3. Wave 4 Older Adult sample

In Wave 4 the CAPS was expanded to include an Older Adult sample. The Older Adult sample consisted of all members of original Wave 1 households who would have been 50 or older on 1 January 2006. This included Older Adults who co-resided with a Young Adult in Wave 1 and Older Adults from the households that included no Young Adults.

3.4. Wave 4 Child sample

In Wave 4 we attempted to interview the primary caregiver and measure the height (or length) and weight of all children born to female young adults who were successfully re-interviewed in Wave 4.

4. Fieldwork and training

4.1. Wave 1 fieldwork and training

The first interviews for CAPS were conducted in early August, 2002. The fieldwork in this first wave comprised the administration of 'young adult' and 'household' questionnaires as well as the evaluation of the literacy and numeracy of the young adults in the sample.

Fieldwork for CAPS Wave 1 was contracted to Markinor, a well-known South African survey research organization based in Johannesburg, under the direction of Anneke Greyling. Markinor employed a team of approximately 120 interviewers, a few of whom conducted over one hundred interviews each (and one of whom conducted 150 interviews). CAPS project staff carried out the training of Markinor field teams in early August 2002, and were involved in quality control throughout the fieldwork. The bulk of the fieldwork was concluded in December, 2002, with some additional interviews conducted in predominantly white areas in early 2003.

The questionnaire was administered as a face-to-face in-home interview using a paper questionnaire. The fieldworkers completed the questionnaires in English but had copies translated into Afrikaans and Xhosa so that the questions could be asked in the language of the respondent. The questionnaires were tested in two rounds of pilot interviews. The literacy and numeracy evaluation was tested in pilots in several schools.

Interviewers were deployed according to the majority racial population of the area. For example, white interviewers were used in predominantly white areas. Young women respondents were supposed to be interviewed by female interviewers, and young men respondents by male interviewers, given the sensitivity of questions about sex and health. In practice, we discovered afterwards, this rule was not applied uniformly. Almost all young white women and most young coloured women were interviewed by female interviewers, but almost one half of young African women were interviewed by male interviewers. This was not our intention, and there is a clear need for further research on the sensitivity of some responses to the gender of the interviewer

Respondents were only interviewed if and when they had signed a consent form that provided them with the information required for informed consent. In the cases of young people below the age of eighteen years, a parent or guardian was also required to sign the consent form for the 'young adult' interview. Interviewees were given a small gift (a strong bag, with a value of about R40 or US$5) as a token of our appreciation.

The protocol for fieldwork stipulated that only households indicated on the aerial photographs as selected households in each EA were to be interviewed. There would be no substitution if the selected household was unavailable. Each household was to be contacted at least five times, with contacts made at different times and different days, including at least two attempts on an evening or weekend.

Field teams had to be pulled out of five EAs which were considered too dangerous to work in. The team was able to return to one of these EAs after mediation created an environment

in which the field work could be continued. The sample weights that adjust for household non-response include adjustments for the loss of EAs in a given population group cluster.

4.2. Wave 2 fieldwork and tracking

The Wave 2a fieldwork was conducted in different areas by two teams, between July and November 2003. Fieldwork in predominantly Coloured areas was conducted by Development Research Africa (DRA). Research in African areas was conducted by a new in-house, Xhosa-speaking fieldwork team recruited and trained within the CSSR, and supervised by Jo Stein. Interviews with white respondents were conducted by students from the University of Cape Town, supervised by Viki Elliott.

Wave 2b (2004) fieldwork in predominantly Coloured and white areas was conducted by Citizen Surveys, whereas fieldwork in African areas was again conducted by the CSSR's own fieldwork team (under the supervision of Viki Elliott). Fieldwork began in April 2004 in predominantly African areas and in July in predominantly white and coloured areas. Most fieldwork was completed in November 2004, with a final few interviews completed in December. In Wave 2b, the remaining two-thirds of the Young Adult sample were re-interviewed. A small number of young adults chosen for the Wave 2a sample, but not successfully interviewed were re-contacted in Wave 2b.

In order to reduce attrition in the panel, we collected a range of contact information in the first wave of CAPS. This included:

- Address, telephone numbers and email addresses (where appropriate), for both the young adult respondents and the adult household member who completed the household questionnaire;
- Names, addresses and telephone numbers of up to three people who know the household well;
- Names, addresses and telephone numbers of up to three people who know the young adult well, outside of the household (there might be an overlap between this list and the previous, household list);
- We also have maps of the EAs with streets marked, showing where the young adult and household lived in Wave 1.

A system of quality control for fieldwork conducted by the UCT team was initiated in Wave 2a and expanded in Wave 2b. All questionnaires were administered as face-to-face interviews using paper questionnaires and data capture was completed by each organization for their own fieldwork.

4.3. Wave 3 fieldwork and training

In March 2005, before Wave 3 fieldwork began, we sent out the first issue of the *CAPS Newsletter*. This newsletter is essential to the task of keeping our respondents informed about the study, which is important ethically (so that respondents can consent on an informed basis to remain in the study) and practically (as regular communication is expected to reduce attrition). Additionally, we included a sheet intended for the respondents to update send back their contact information, which was successful as we received many sheets back. Often, a family member at the old residence would forward on the contact sheet to the young adult.

Fieldwork for Wave 3 of CAPS was conducted between April and December 2005. As in Wave 2b, the fieldwork in predominantly white and Coloured areas was conducted by Citizens Surveys, and the fieldwork in predominantly African areas conducted by the CSSR/UCT fieldwork team under the direction of Viki Elliott.

The target sample for Wave 3 was the full set of 4,752 young adults originally interviewed in Wave 1, except those known to be deceased or mentally ill from Wave 2 fieldwork. This resulted in a starting target of 4,737 Young Adults. Fieldworkers were instructed to first attempt to contact the Young Adult, either by telephone or by visiting their most recent address. If this did not prove successful, fieldworkers contacted the three contact people, given by the respondent in Wave 1 and updated in Wave 2.

Approximately 80% through the Wave 3 fieldwork timeline, a separate tracking team of in-house fieldworkers was designated. These fieldworkers re-contacted target respondents with incomplete interviews and those that had indicated they would be available later in the year. Also, fieldworkers visited contact people without telephone numbers on foot. This tracking operation was highly successful; however, no interviews were conducted with respondents who had moved outside of metropolitan Cape Town.

At every household where a young adult was successfully interviewed, interviewers were instructed to administer a household questionnaire to a household member aged 18 or over and knowledgeable about the household's members and finances. In addition, a short 'parent' questionnaire was administered to the parents or guardians of our young adult respondents, probing their attitudes and beliefs on education and value socialization, their assessment of the home environment in which the respondents had grown up, and their social and economic expectations for and of their children. This questionnaire was only administered to co-residential parents (covering approximately one third of our sample of young adults), for practical reasons. Parent or guardians were identified from the young adults' Wave 1-2 households. The target sample of "parental figures" was limited to the following types of relationship to the young adult: father/mother; stepfather/mother; adoptive/foster parent; grandparent, and uncle/aunt. Approximately 2000 successful parent interviews were completed.

Quality control for Wave 3 was completed by a team at UCT for UCT fieldwork, and at Citizens Surveys for their fieldwork, but was synchronized and some Citizens Surveys fieldwork was also checked by UCT quality control for consistency. The data capture program was written with substantial input from the UCT team, and all data capture was performed by Citizens Surveys.

4.4. Wave 4 fieldwork and training

Fieldwork for Wave 4 of CAPS was conducted between April and December 2006. As in Waves 2b and 3, the fieldwork in predominantly white and Coloured areas was conducted by Citizens Surveys, and the fieldwork in predominantly African areas conducted by the SALDRU/UCT fieldwork team under the direction of Lebo Sello and Viki Elliott. Initial training of the UCT field team was done by Anne Case, David Lam, Murray Leibbrandt, Alicia Menendez and Cally Ardington. Fieldworkers then had an additional week of training in small groups under the supervision of Lebo Sello with the assistance of Thobani Ncapai,

Bulelwa Nokwe and Nobulele Mata. Citizen Surveys fieldworkers were trained by David Lam, Murray Leibbrandt and Lebo Sello.

There were three target samples for Wave 4. The Young Adult sample was comprised of the full set of 4,752 young adults originally interviewed in Wave 1, except those known to be deceased or mentally ill from previous waves. This resulted in a starting target of 4701 Young Adults. The Older Adult sample consisted of all members of original Wave 1 households who would have been 50 or older on 1 January 2006. This included Older Adults who co-resided with a Young Adult in Wave 1 and Older Adults from the households that included no Young Adults. The third target sample was all children born to female young adults.

At every household where a young adult, older adult or child was successfully interviewed, interviewers were instructed to administer a household questionnaire to a household member aged 18 or over and knowledgeable about the household's members and finances.

Similar to Wave 3, quality control for Wave 4 was completed by a team at UCT for UCT fieldwork, and at Citizens Surveys for their fieldwork, but was synchronized and some Citizens Surveys fieldwork was also checked by UCT quality control for consistency. The data capture program was written by a UCT team, and all data capture was performed by UCT. Every questionnaire was double captured and a full reconciliation of conflicting answers was performed.

In March and May of 2007 we made our first attempts to track respondents outside of metropolitan Cape Town with two field trips to the Eastern Cape. We successfully located and interviewed 31 young adults and 17 older adults.

5. Non-response and attrition

5.1. Wave 1 sample and response rates

5.1.1. Profile of the CAPS Wave 1 sample

Table 1 provides an overview of the CAPS sample by the predominant population group of the enumeration area, based on the records in the household and young adult files. As explained in the section on sample design, EAs were classified according to the predominant population group of household heads in the EA. As shown in Table 1, there are 22,631 individual records in the household file, representing 5,256 households. There are 4,752 young adults aged 14-22 in the young adult file. Broken down by the classification of enumeration areas, 42% of the individuals in the household sample came from EAs that were classified as African, 44% came from EAs classified as coloured, and 14% came from EAs classified as white. These percentages intentionally do not correspond to the actual distribution by population group in Cape Town, since the goal was to produce a larger number of African young adult respondents than would be found in a random sample. Sample weights are necessary to adjust the sample back to proportions that are closer to the actual population of Cape Town.

Table 2 breaks down the CAPS sample by the actual population group of respondents. Taking the population group as reported for each household member by the household respondent, 42% of the household sample is African, 46% is coloured, and 11% is white. It is noteworthy that the percentage of African household members in Table 2 is very similar to the percentage of household members from African EAs in Table 1, while the percentages of coloured and white household members differ from the percentages from coloured and white EAs. The percentage of household members classified as coloured in Table 2 is higher than the percentage of household members from coloured EAs in Table 1, while the percentage of household members

Table 1. Numbers of Household Members, Households, and Young Adults in CAPS Sample, by Predominant Population Group of EA

Population Group of Enumeration Area	*Household members*		*Households*		*Young adults*	
	Number	*%*	*Number*	*%*	*Number*	*%*
African	9,565	42.3	2,260	43.0	2,126	44.7
Coloured	9,884	43.7	2,036	38.7	1,879	39.5
White	3,182	14.1	960	18.3	747	15.7
Total	22,631	100.0	5,256	100.0	4,752	100.0

classified as white is lower than the percentage of household members from white EAs. The explanation for this difference is fairly simple. EAs classified as white are the least homogenous of the three types of EAs. In the group of EAs with predominantly African household heads (that is, the group we classify as African EAs), 98% of the households are African. In the EAs we classify as coloured, 96% of the households are coloured. But in the

EAs we classify as white, only 87% of households are actually white, with 10% being coloured and 3% being African. This means that our so-called "white" EAs produced a sample that included a significant number of coloured households. This was an issue that we took into account in our sample design, but it is possible that EAs became even less homogenous between 1996 and 2001. Since Statistics South Africa had not made EA-level data available from the 2001 census, it was impossible to directly test the extent to which the racial composition of EAs changed between the 1996 census and the time of our survey.

Table 2. Numbers of Household Members, Households, and Young Adults in CAPS Sample, by Population Group of Respondent

Population group	*Household members*		*Households**		*Young adults*	
	Number	*%*	*Number*	*%*	*Number*	*%*
African	9,540	42.1	2,275	43.3	2,144	45.1
Coloured	10,419	46.0	2,169	41.3	1,976	41.6
Indian	99	0.4	19	0.4	22	0.5
White	2,426	10.7	772	14.7	593	12.5
Other/missing	153	0.7	21	0.4	17	0.4
Total	22,631	100.0	5,256	100.0	4,752	100.0

**Population group of person with person code=1 is used for household*

Taking the self-classification of population group by young adults, the young adult sample consists of 2,144 African YAs, 1,976 coloured YAs, and 593 white YAs. Our goal of producing roughly equal numbers of African and coloured respondents was very successful. Our goal of producing a number of white respondents that was roughly half the number of African and coloured respondents was less successful. This was a result of several factors. The first factor is the issue just discussed about the relative heterogeneity of EAs classified as white. This meant that our sample design produced fewer white households than projected. The second important factor is that our response rates in white areas were significantly lower than the response rates in African and coloured areas, a result consistent with the experience of virtually all household surveys in South Africa. This was also anticipated, and our response rates were roughly in line with our expectations. Our sample of 593 white YAs is large enough for statistically meaningful analysis of white young adults in Cape Town, and is considerably larger than the sample of white young adults in Cape Town than is available in national surveys. The Labour Force Survey, for example, has roughly 200 whites aged 14-22 in the Western Cape in a given round of the survey.

As discussed above in the section on sample design, we selected up to three young adults per household into the young adult sample. Table 3 shows the numbers of young adults per household in the final YA sample, looking at the sample with completed questionnaires. Looking at the final column, roughly 45% of the YAs came from households with only one YA, 38% came from households with two YAs, and 17% came from households with three YAs. The distribution differs across population groups. In the African sample, 41% of YAs came from households with only one YA, compared to 53% in the white sample. The number of households represented in the young adult file (counting only young adults with completed interviews) is 3,304, with 1,435 African households, 1,395 coloured households, and 445 white households.

Table 3. Numbers of Young Adults per Household in CAPS Sample

Number of YAS in Household	*African*		*Coloured*		*White*		*Indian/Other*		*Total*	
	N	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*
1	883	41.2	908	46.0	316	53.3	21	53.8	2,127	44.8
2	790	36.8	785	39.7	223	37.6	13	33.3	1,812	38.1
3	471	22.0	283	14.3	54	9.1	5	12.8	813	17.1
Total	2,144	100.0	1,976	100.0	593	100.0	39	100.0	4,752	100.0

The large number of YAs who come from households with more than one YA should have important payoffs to researchers. The fact that over half of our sample of young adults has at least one other household member in the sample means that CAPS will be a valuable resource for analysis of issues related to intra-household allocation of resources. Looking at all of the individuals age 14-22 in the household file, 2.2% were excluded from the YA sample by our condition that no more than three were selected per household.

5.1.2. Wave 1 Response Rates

Response rates for CAPS Wave 1 can be calculated for both the household sample and the young adult sample, since issues of contact and refusal come up first in the creation of the household sample and subsequently in the attempt to interview the young adults who are selected from each household. Key pieces of information for the calculation of response rates are presented in Table 4. As shown in Row 1 of Table 4, the total number of dwellings screened for the CAPS sample was 11,561. This includes all dwellings selected using the sampling procedure outlined above, plus secondary households added during the screening process. There were 265 dwellings that were unoccupied or were not residential households, bringing the number of screened occupied households to 11,296. Broken down by the majority population group in the enumeration area, there were roughly 3,500 screener households in African areas, 3,400 screener households in coloured areas, and 4,500 screener households in white areas. It is worth noting that we screened 1,000 more households in white areas than in African areas, even though our target number of white young adults was only half of our target number of African young adults. This is a reflection of how difficult it is to locate a sample of white young adults in Cape Town (or indeed anywhere in South Africa). There are three major factors explaining why a much larger screener sample is required to locate a given number of white young adults than to locate the same number of African or coloured young adults. First, as noted above, white households are much less likely to have 14-22 year-old residents, a reflection of the older age structure and smaller family size among whites. Second, neighbourhoods that are predominantly white are more heterogeneous in terms of population group than neighbourhoods that are predominantly African or coloured, making it harder to target white households. Third, expected response rates are much lower in white areas, requiring a larger initial sample in order to achieve a target number of successfully interviewed households.

Table 4. Estimated Response Rates in CAPS Wave 1

	Majority Population Group of Enumeration Area			*Total*
	African	*Coloured*	*White*	
1. Number of dwellings screened	3,590	3,429	4,542	11,561
2. Number of occupied households screened	3,518	3,374	4,404	11,296
3. Percent of households with completed screeners	90.5%	85.3%	59.0%	76.7%
4. Estimated number of selected households	2,534	2,457	2,089	7,080
5. Completed household interviews	2,260	2,036	960	5,256
6. Household response rate (line 5/ line 4)	89.2%	82.9%	46.0%	74.2%

Among the 11,296 screened households, our sample design called for us to select all of the households that had 14-22 year-old residents in the household, 48% of the remaining households in African and coloured areas, and 28% of the remaining households in white areas. Using our projection that 50% of African and coloured households and 25% of white households would have young adults, and given our selection rule for remaining households, a simple projection of the number of selected households that would be produced by from our 11,296 screener households would be 7,125 selected households.

The actual number of selected households cannot be calculated precisely because some households were not available or refused to answer the screener questionnaire. The percentage of households with completed screener questionnaires is shown in Line 3 of Table 4. Over 90% of screener questionnaires were completed in African areas, compared to only 59% in white areas, with 77% completed in all areas combined. Table 5 shows the distribution of response codes in each EA group. About 26% of white households were coded as "not available," meaning that field teams were unable to make contact with anyone in the household after at least five attempts. In many cases these were households in gated security complexes where it was impossible to get access to any households in the complex. Cases of outright refusal were 2.8% in African areas, 6% in coloured areas, and 15% in white areas.

Table 5. Response Codes for Household Screening Questionnaire in CAPS Wave 1

Response Code	*Majority Population Group of Enumeration Area*			*Total*
	African	*Coloured*	*White*	
Completed	90.5%	85.3%	59.0%	76.7%
Not Available	6.7%	8.7%	26.1%	14.9%
Refused	2.8%	6.0%	14.9%	8.5%
Total	100%	100%	100%	100%

For households without completed screeners it is impossible to know whether the households contained young adults, and therefore it is impossible to know with certainty whether they would have been selected into the sample. A good estimate can be made, however, by using the fact that some fraction of households would have been selected on the basis of our screening number rule, whether or not they contained young adults. For households that did

not pass the screening number rule, we can use the probability of having a 14-22 year old household member that applies to the population group area of the household's EA. For example, if the household was located in an EA classified as white, we assume that there was a 25% probability that the household contained a 14-22 year-old household member. Taking these estimates for the 23% of households without screeners and combining them with the actual number of selected households for the 77% of households with completed screeners, we produce an estimated number of selected household for each group of EAs.

As shown in Line 4 of Table 4, the estimated number of households selected into the CAPS sample is 7,080. If our response rate had been 100%, this would have been the number of completed household interviews in the final sample. As shown in Line 5, the actual number of completed household questionnaires is 5,256. As shown in Line 6, this is a response rate of 74% for all EA groups combined. The rate varies dramatically across the three EA groups, with a response rate of 89% in African areas, 83% in coloured areas, and 46% in white areas. The low response rate in white areas is unfortunately a common result for household surveys in South Africa. As was shown in Table 5, a large contributing factor is the difficulty in securing access to households in these areas, with concerns about security making it very difficult to even make contact with many households. Advance letters and negotiations with neighbourhood committees helped gain access in many areas, but many areas remained impossible to reach. Table 5 also shows that outright refusal rates are also a problem in white areas, even when contact can be made with the household. While we believe that the sample of white households is an important component of the CAPS sample, the roughly 50% response rate in white areas means that researchers should be cautious in drawing inferences about the white population.

Although response rates vary dramatically across the three EA groups there is also considerable variation in EA level response within each of the groups. Table 6 presents results from an analysis of enumerator area response rates. The response rate for each enumerator area is regressed on the majority population group and the logarithm of the average household income in the enumerator area from the 2001 Census. The first column confirms results from tables 4 and 5 in showing response rates to be much lower in enumerator areas where the majority population group is zero. In the second column average household income is added to the regression with the result that the African and coloured dummies are reduced by a half and a third respectively. There is a significant negative relationship between average household income and the EA level response rate. This is true within each of the three EA groups. It appears from Table 6 that much of the differential in response rates across the three EA groups can be attributed to differences in household income.

Table 6: EA level response rates: coefficients and standard errors from OLS regression of EA level response rates on majority population group and average household income.

African	0.439	[0.022]**	0.202	[0.042]**
Coloured	0.365	[0.022]**	0.245	[0.028]**
Logarithm of average household income			-0.114	[0.018]**
Constant	0.454	[0.014]**	1.825	[0.214]**
Observations	405		405	
R-squared	0.55		0.59	
Standard errors in brackets				
* significant at 5%; ** significant at 1%				

5.1.2.1. Wave 1 Response Rates for Young Adults

In addition to the issue of response rates at the household level, we must also consider response rates among young adults who were selected to be in the young adult sample. Once young adults had been selected into the YA sample based on the household roster in the household questionnaire, field teams were not always able to successfully make contact with the young adults. Young adults also had the option of refusing to answer the YA questionnaire. Table 7 shows the key pieces of information we used to calculate response rates for young adults. As shown in Line 1 of Table 7, we estimate that 4,637 households with young adults would have been selected into the sample. As discussed above, this estimate is based on the assumption that households that did not complete the screener questionnaire had the same probability of having young adults as other households in the same population group cluster. The actual number of households containing young adults with completed questionnaires was 3,493. This yields a household response rate for households with young adults of 75.3%, very similar to the household response rate for all households. The response rate is 89% for African households, 82% for coloured households, and 48% for white households.

Line 4 of Table 7 shows that 5,271 young adults were selected into the young adult sample based on the household roster. Line 5 shows that we ended up with 4,752 completed young adult questionnaires. As shown in Line 6, this is a response rate of 89.6% for young adults conditional on the household having been interviewed. Differences across population groups are much smaller than they were for overall household participation. Response rates for young adults were 93% for African young adults, 88% for coloured young adults, and 86% for white young adults. The net response rate for young adults is the product of the probability that their household was successfully interviewed times the probability that the young adult was successfully interviewed. As shown in Line 7 of Table 7, this composite young adult response rate is 83% for African young adults, 72% for coloured young adults, and 42% for white young adults.

Table 7. Estimated Response Rates for Young Adults, CAPS Wave 1

	Majority Population Group of Enumeration Area			*Total*
	African	*Coloured*	*White*	
1. Estimated number of selected households with young adults	1,657	1,713	1,267	4,637
2. Completed household interviews in households with young adults	1,472	1,410	611	3,493
3. Household response rate for households with young adults (line 2/ line 1)	88.8%	82.3%	48.2%	75.3%
4. Number of young adults selected	2,285	2,148	869	5,302
5. Total completed young adult interviews	2,126	1,879	747	4,752
6. Young adult response rate (conditional on household response) (line 5/line 4)	93.0%	87.5%	86.0%	89.6%
7. Composite young adult response rate (line 6 x line 3)	82.6%	72.0%	41.5%	67.5%

Because information on young adults is provided in the household questionnaire, we have a good deal of information about the roughly 10% of young adults that we did not succeed in interviewing. Using this information, we can analyze the characteristics of non-respondents compared to respondents. Table 8 presents young adult response rates by age, gender, and population group, based on the information provided about the young adults in the household questionnaire. Not surprisingly, Table 8 shows a strong relationship between response rates and age. Looking at the last column, which combines males and females for all population groups, response rates for young adults age 20 and above were 85% to 87%, while response rates for those under age 18 were 93% to 95%. Looking at the separate groups by age, gender, and population group, the highest response rates are for young African females, who had response rates of 96%-99%, while the lowest response rates are for white and coloured males, who had response rates as low as 70%. Response rates are generally higher among Africans, with response rates averaging 96% for females and 92% for males. While the lowest response rates are among whites, a result that was expected, it is noteworthy that response rates among coloured males are very similar to those for white males, around 83%. Response rates among white females are around 90%, similar to the overall average.

Table 9 shows the percentage of young adults who refused to answer the young adult questionnaire, using the same breakdown by age, gender, and population group. Comparing Table 8 with Table 7, it appears that the high rates of non-response among older coloured males appear to be more the result of non-contact rather than refusal. Non-response among older white males, on the other hand, appears to be more driven by refusals.

Table 8. Percentage of Selected Young Adults who Completed YA Questionnaire by Age, Gender, and Population Group

Population Group	*African*		*Coloured*		*White*		
	Female	*Male*	*Female*	*Male*	*Female*	*Male*	*Total*
14	95.9	94.2	97.3	90.4	90.0	94.9	94.1
15	99.2	96.1	95.5	89.3	97.6	92.3	94.9
16	98.7	91.8	98.0	86.6	81.8	87.2	93.1
17	96.6	95.0	95.1	88.8	97.8	82.9	93.2
18	94.0	87.9	91.7	86.2	84.1	78.7	89.0
19	99.4	92.8	92.9	81.3	85.4	81.8	91.0
20	91.8	88.8	89.7	77.8	85.7	69.0	86.3
21	94.6	92.7	88.5	74.0	92.5	71.4	87.3
22	92.5	90.3	82.8	69.9	89.7	84.0	85.3
Total	95.8	92.1	92.7	83.4	89.7	82.9	90.5

Table 9. Percent Refusing to Answer YA Questionnaire by Age, Gender, and Population Group

Population Group	*African*		*Coloured*		*White*		
	Female	*Male*	*Female*	*Male*	*Female*	*Male*	*Total*
14	0.0	0.8	0.9	4.3	7.5	2.6	2.1
15	0.8	1.9	0.7	1.5	0.0	7.7	1.6
16	1.3	1.0	0.7	3.1	12.1	6.4	2.5
17	2.1	3.0	2.8	4.3	0.0	14.6	3.6
18	3.6	5.6	2.8	1.6	6.8	8.5	4.0
19	0.0	1.4	2.4	4.9	9.8	11.4	3.2
20	2.5	4.3	6.0	8.7	5.7	17.2	5.8
21	2.3	1.8	3.8	8.0	2.5	22.9	5.0
22	2.3	0.9	12.1	3.2	2.6	8.0	4.4
Total	1.7	2.3	3.3	4.4	5.0	10.7	3.6

The high rate of non-contact among older coloured males raises the concern that young adults with jobs may have been disproportionately missed in the fieldwork. Since the household questionnaire includes a question about work activity, we can analyze this issue directly. Tables 10 and 11 compare the response codes for those who were working versus not working for each population group, looking only at those who were age 19-22. Table 10 shows that for African and coloured males there are relatively large differences in non-response between working and non-working respondents. Among African males over 17% of those who were working were never successfully contacted, compared to only 3% of those who were not working. Among white males there is very little difference in the "not available" percentage for those working versus non-working. Surprisingly, there is a higher refusal rate among non-working white males than among working white males.

Table 11 shows the same breakdown for females. In addition to overall lower non-response rates for females compared to males, there appears to be less effect of working on female response rates. There are somewhat higher non-contact rates for those who are working among white and coloured females, but very little difference for Africans.

Table 10. Response Code on YA Questionnaire by Work Status, Males Age 19-22

Response Code	*African*		*Coloured*		*White*		
	Not working	*Working*	*Not working*	*Working*	*Not working*	*Working*	*Total*
Completed	93.8	76.5	79.8	72.7	74.1	77.9	82.0
Not available	3.0	17.4	12.7	18.4	7.4	7.8	10.6
Refused	2.4	3.1	5.2	7.4	18.5	13.0	5.9
Other	0.9	3.1	2.3	1.6	0.0	1.3	1.5
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 11. Response Code on YA Questionnaire by Work Status, Females Age 19-22

Response Code	*African*		*Coloured*		*White*		*Total*
	Not working	*Working*	*Not working*	*Working*	*Not working*	*Working*	
Completed	94.3	93.9	91.0	86.3	83.1	91.3	91.2
Not available	3.3	3.0	2.7	7.3	10.8	1.3	4.2
Refused	1.8	3.0	4.9	5.9	4.6	6.3	3.8
Other	0.7	0.0	1.4	0.5	1.5	1.3	0.8
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Additional analysis of non-response in the young adult sample is presented in the Table 12. The table presents results of probit regressions in which a variable indicating non-response is regressed on a number of individual and household characteristics, including age, gender, population group, work status, years of completed education and quintiles of household income per capita. These regressions reconfirm in a multivariate context the results shown above – non-respondents in the young adult sample are more likely to be older, male, working, and non-African. There appears to be very little relationship between the probability of non-response and household income, however, once we control for the other variables included in the regression.

Table 12. Probit Regression for non-response in CAPS Wave 1 Young Adult Sample

Female	-0.062	[0.008]**
African	-0.095	[0.014]**
coloured	-0.043	[0.013]**
Age 15	-0.009	[0.020]
Age 16	0.032	[0.025]
Age 17	0.032	[0.024]
Age 18	0.112	[0.032]**
Age 19	0.09	[0.031]**
Age 20	0.155	[0.037]**
Age 21	0.138	[0.037]**
Age 22	0.166	[0.040]**
Years of completed education	-0.009	[0.002]**
Working	0.033	[0.012]**
Per capita income quintile 1	0.009	[0.014]
Per capita income quintile 2	0.038	[0.016]*
Per capita income quintile 3	-0.002	[0.015]
Per capita income quintile 4	-0.004	[0.017]
Income missing	0.042	[0.023]
Observations	4691	

Standard errors in brackets

* significant at 5%; ** significant at 1%

5.2. Waves 2, 3 and 4 young adult response and attrition

Young adults were identified as part of the CAPS panel if they successfully completed a Wave 1 interview. In subsequent waves, we made no attempt to re-contact individuals identified as young adults but not successfully interviewed in Wave 1. Contact information collected in Wave 1 proved invaluable in tracking respondents in subsequent waves, especially when respondents moved but also when they stayed at an address that was difficult to locate. In Waves 2a, 2b and 3 we made no attempt to re-interview members of the panel who had moved outside of Cape Town. In Wave 4 we tracked a sample of young adults who had moved to the Eastern Cape. We conducted 31 Wave 4 interviews with young adults in the Eastern Cape.

An unfortunate reality in panel surveys is attrition across waves as the survey progresses. Table 13 shows the response rate for each of Waves 2, 3 and 4 by population group of the young adults. Response rates are calculated as a percentage of the successful Wave 1 interviews even though we did not attempt to re-interview those young adults known from previous fieldwork to be deceased or mentally ill. Overall response rates were 83%, 74% and 72% in Waves 2, 3 and 4 respectively. In addition to poor initial response rates it is clear from Table 13 that attrition is the most serious in the white group. In Wave 4 we managed to re-interview 74% and 80% of our original African and coloured young adults as opposed to only 42% of white young adults. High non-response among the white population is a common issue facing surveys in South Africa. The 1993 Project for Statistics on Living Standards and Development (PSLSD) conducted by the Southern Africa Labour Development Research Unit at the University of Cape Town found systematically higher non-response among whites, who were more likely to refuse the interview than members of other population groups (SALDRU 1994). Similarly, the 2005 South African National HIV Prevalence, HIV Incidence, Behaviour and Communication Survey, conducted by the Human Sciences Research Council (HSRC), the Centre for AIDS Development, Research and Evaluation (CADRE) and the Medical Research Council (MRC) also reports higher non-response for white households than for African or Coloured households (Shisana et al, 2005). They report a response rate of 67.6% for white households, compared to 88.5% for Africans and 87.8% for coloured households. Furthermore, they also find the lowest response rates in urban formal areas compared to rural formal areas, so the white urban rate (more comparable to CAPS' white sample) is potentially even lower than the 67.6% reported for white households overall.

Table 13. Response rates across waves 2, 3 and 4

	African	*Coloured*	*White*	*Total*
Wave 1 successful interviews	2151	2005	596	4752
Wave 2 successful interviews	1821	1693	413	3927
Wave 2 response rate	84.7%	84.4%	69.3%	82.6%
Wave 3 successful interviews	1515	1679	337	3531
Wave 3 response rate	70.4%	83.7%	56.5%	74.3%
Wave 4 successful interviews	1596	1594	249	3439
Wave 4 response rate	74.2%	79.5%	41.8%	72.4%

Table 14 shows a break-down of the components of non-response for each wave. Non-response is recorded in the following categories: not available (often a soft refusal), deceased, refused, moved, institutionalised (jail, hospital or rehabilitation centre), and no contact. Moves were captured during fieldwork into the four different categories. The first category of moved and not successfully interviewed is "moved within Cape Town". In these cases, an interviewer was able to contact someone such as a neighbour who knew the respondent had moved from previous address but stayed within Cape Town. However, this person was unable to give a phone number, new address, or any details leading to a successful interview. The second and third categories are "moved within South African" and "moved abroad". Here, the interviewer made contact with someone, including potentially the young adult themselves via phone, who knew that the young adult had moved out of Cape Town, either within South Africa or to another country. The final classification, "moved no details", is used when the interviewer was able to contact someone who knew the respondent had moved from previous address, but did not have any further details.

The most common causes of non-response vary by population group and wave. Overall, the most common type of non-response recorded for Wave 2 was "not available", which is generally a "soft refusal" by the respondent. However, for African young adults, "moved out of range" is a more frequent cause of non-response. These respondents had moved mostly to the Eastern Cape, reflecting the migratory cycles of young African people between the Western and Eastern Cape. Moving out of Cape Town was again a common cause of non-response for white young adults, second to "unavailable". However, white young adults are more likely to leave South Africa and head overseas, particularly to the United Kingdom. By contrast, a smaller percentage of Coloured young adults moved outside of Cape Town, reflecting the closer ties of the Coloured population to the Cape Town area. A very small number (less than 4 percent of the original panel) had died or were in prison, hospital or a rehabilitation centre.

In Wave 3, the most common reason for non-response is moving out of Cape Town. There is also a substantial proportion of non-response that is due to respondents moving to a location about which we have no information (moved no details). Refusal rates amongst Africans remain substantially lower than those for coloureds and whites. African respondents are still most likely to move within South Africa. Coloured non-response is primarily driven by hard and soft refusals and moving, either within South Africa or to an unknown destination. Almost half of white non-response is accounted for by hard and soft refusals. A substantial proportion of whites have moved abroad or to an unknown destination.

The most common reason for non-response in Wave 4 is respondents moving to an unknown destination. The increase in this category across the waves, particularly for Africans, is a reflection of the difficulty in tracking people over time, particularly those living in more informal areas. Coloured and white non-response is once again driven by hard and soft refusals.

Table 14: Break-down of non-response across waves 2, 3 and 4

	Wave 2			
Reason for non-response	*African*	*Coloured*	*White*	*Total*
Not Available	30.9	45.2	38.8	37.9
Refused	8.2	19.9	29	17.2
Deceased	3.3	1.6	0	1.9
Moved within Cape Town	7.6	3.2	0	4.2
Moved within SA	42.4	16	12.6	25.8
Moved abroad	0.3	5.8	17.5	6.2
Moved no details	3.3	4.2	0.6	3
Institutionalised	1.2	3.2	0	1.7
Mentally Unfit/Disabled	0	0.6	0	0.2
No contact	2.7	0.3	1.6	1.6
	Wave 3			
Reason for non-response	African	Coloured	White	Total
Not Available	16.2	17.2	19.7	17.2
Refused	7.2	21.8	28.6	15.6
Deceased	5.7	2.8	0.4	3.8
Moved within Cape Town	16.2	8.6	1.2	11
Moved within SA	34.1	14.7	13.1	24.5
Moved abroad	0.2	5.8	18.2	5.5
Moved no details	17.6	23.3	17	19
Institutionalised	2.5	3.4	0.8	2.4
Mentally Unfit/Disabled	0.3	0.6	0	0.3
No contact	0	1.8	1.2	0.7
	Wave 4			
Reason for non-response	African	Coloured	White	Total
Not Available	5.6	20.9	24.5	15.4
Refused	11.9	26	37.8	23.2
Deceased	7.6	3.2	0.6	4.3
Moved within Cape Town	3.8	1	0.3	2
Moved within SA	30.8	10	6.9	18
Moved abroad	0.2	2.9	9.5	3.5
Moved no details	32.4	27.7	17.3	27
Institutionalised	2.2	2.9	0	1.8
Mentally Unfit/Disabled	0.5	0.7	0.3	0.5
No contact	5.1	4.6	2.9	4.3

The relationship between age and non-response is clearly illustrated in Figure 2. Figure 2 shows response rates for each of the waves by age at Wave 1. The differential in Wave 2 response rates between the youngest and oldest groups is greater than 20 percentage points.

Figure 2: Response rates by wave and age at Wave 1

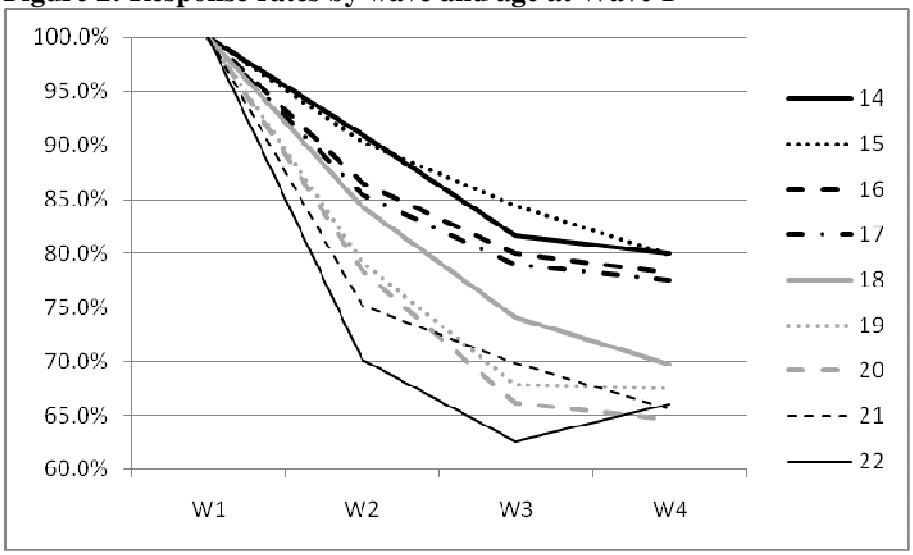


Table 15 examines the determinants of non-response in each wave in a multivariate context. The table present marginal effects and standard errors from probit regressions for each of Wave 2, 3 and 4 in which the dependent variable is set equal to one if the young adult did not successfully complete a young adult questionnaire in that wave, and is set equal to zero otherwise. This non-response variable is regressed on dummy variables for single years of age, gender, population group, school enrolment, employment status, quintiles of per capita income and adding place of birth, household composition and physical household characteristics. All characteristics are taken from the Wave 1 young adult and household data. In addition to the income quintiles a dummy is included for missing income. The marginal effects can be interpreted as the increase (decrease) in the probability of non-response from being in that category, controlling for the other characteristics. For example, in Wave 2 coloured young adults are estimated to have response rates that are 8.5% higher than white young adults, evaluated at the mean values of the other characteristics.

In all three waves, we observe higher non-response amongst older young adults, those not in school, and those born outside of Cape Town. Response rates were consistently lowest amongst white young adults. In Wave 2 coloured non-response was higher than African non-response but this was reversed in Wave 3. Response rates for Africans and coloureds were very similar in Wave 4.

Variables included to represent family and household composition in Wave 1 suggest that young adults with more familial ties to their Wave 1 households were more likely to be successfully re-interviewed in subsequent waves. Living with biological mother has a sizable effect; these respondents were between 7% and 11% more likely to be successfully re-contacted than respondents not living with at their biological mother.

Table 15: Marginal effects and standard errors from probit regressions for Young Adult non-response in CAPS Waves 2, 3 and 4.

	Wave 2	Wave 3	Wave 4
Female	0.016 [0.011]	0.012 [0.013]	0 [0.014]
African	-0.145 [0.022]**	-0.134 [0.027]**	-0.247 [0.027]**
Coloured	-0.085 [0.018]**	-0.192 [0.022]**	-0.248 [0.022]**
Age 15	0.023 [0.029]	-0.033 [0.028]	-0.002 [0.030]
Age 16	0.082 [0.033]*	0.016 [0.030]	0.021 [0.032]
Age 17	0.096 [0.034]**	0.02 [0.031]	0.016 [0.032]
Age 18	0.115 [0.036]**	0.066 [0.034]	0.101 [0.037]**
Age 19	0.171 [0.041]**	0.105 [0.038]**	0.095 [0.038]*
Age 20	0.172 [0.044]**	0.114 [0.041]**	0.122 [0.042]**
Age 21	0.204 [0.046]**	0.072 [0.040]	0.102 [0.043]*
Age 22	0.244 [0.049]**	0.106 [0.044]*	0.055 [0.043]
In School	-0.036 [0.015]*	-0.041 [0.018]*	-0.021 [0.018]
Working	-0.001 [0.015]	-0.032 [0.018]	-0.011 [0.019]
Years of Education	-0.009 [0.004]*	-0.002 [0.005]	0 [0.005]
Matric	-0.03 [0.017]	-0.009 [0.021]	0.016 [0.023]
Eastern Cape	0.081 [0.021]**	0.067 [0.021]**	0.062 [0.023]**
Other South Africa	0.119 [0.023]**	0.099 [0.025]**	0.057 [0.025]*
Outside South Africa	0.294 [0.072]**	0.314 [0.075]**	0.239 [0.076]**
Co-Resident YA	-0.033 [0.011]**	-0.047 [0.013]**	-0.043 [0.013]**
Biological Mother	-0.069 [0.016]**	-0.096 [0.019]**	-0.112 [0.019]**
Biological Father	-0.029 [0.024]	-0.07 [0.029]*	-0.087 [0.030]**
Biological Mother and Father	0.002 [0.028]	0.018 [0.034]	0.026 [0.035]
Indoor Toilet	0.014 [0.021]	-0.123 [0.030]**	-0.048 [0.030]
Electricity	-0.098 [0.029]**	-0.089 [0.030]**	-0.085 [0.032]**
Indoor Water Source	0.028 [0.017]	0.027 [0.020]	0.048 [0.020]*
Quintile 2	-0.003 [0.017]	0.027 [0.020]	0.03 [0.021]
Quintile 3	0.016 [0.019]	0.023 [0.022]	0.034 [0.023]
Quintile 4	0.014 [0.020]	0.022 [0.024]	0.033 [0.025]
Quintile 5	0.053 [0.026]*	0.113 [0.031]**	0.14 [0.032]**
Missing Income	0.05 [0.029]	0.106 [0.035]**	0.153 [0.036]**
Observations	4693	4679	4648

Standard errors in parentheses

* significant at 5%; ** significant at 1%

A selection of variables was included in both regressions to represent the type of housing of the young adults in 2002. Being connected to an electricity supply had a significant and positive effect on the likelihood of a young adult being re-interviewed. Response rates however are not significantly different for young adults with/out piped water and having an indoor toilet is only a significant predictor of non-response in Wave 3.

While young adults in the second, third and fourth quintiles have slightly higher non-response rates than those from the poorest quintile, these differences in response rates are not

significant. The respondents in the richest quintile and those with missing income have significantly higher non-response than those in the poorest quintile.

5.3. Wave 3 and 4 household response rates for young adult households

In Waves 3 and 4, a separate household questionnaire was administered at households where there was a completed Wave 3 or 4 Young adult questionnaire. As a result, except for a few cases where a household questionnaire was not completed at the household of a completed young adult, non-response of households is linked to young adult non-response. Table 16 shows the percentage of completed household questionnaires. In Wave 3 and 4, 4.6% and 2.1% of young adult questionnaires are not accompanied by a completed household questionnaire. Many of these cases are due to interviewee fatigue.

Table 16. Characteristics of Wave 3 and Wave 4 Household Questionnaire respondents

	Wave 3				*Wave 4*			
	African	*Coloured*	*White*	*Total*	*African*	*Coloured*	*White*	*Total*
Completed YAs with:								
Complete Household Q	95.91%	96.96%	95.85%	96.40%	99.00%	96.68%	98.80%	97.91%
Characteristics of Household respondent								
Gender:								
Male	32.7%	17.4%	30.6%	25.0%	20.82%	23.67%	33.89%	23.29%
Female	67.3%	82.6%	69.4%	75.1%	79.18%	76.33%	66.11%	76.71%
Working:								
Yes	34.82%	49.28%	72.16%	45.74%	49.3%	40.99%	61.09%	46.3%
No	65.18%	50.56%	27.84%	54.18%	50.7%	59.01%	38.91%	53.7%
Respondent is a YA:								
Yes	68.18%	14.26%	21.57%	36.88%	26.16%	30.13%	32.64%	28.6%
No	31.82%	85.74%	78.43%	63.12%	73.84%	69.87%	67.36%	71.4%
Relationship to head of household:								
Self	26.0%	46.4%	47.5%	38.2%	56.0%	41.5%	38.9%	47.6%
Spouse/partner	12.2%	37.2%	42.0%	27.6%	23.8%	25.1%	32.2%	25.1%
Biological son/daughter	39.3%	9.8%	7.1%	21.5%	9.0%	22.4%	23.9%	16.6%
Sibling	6.7%	1.1%	1.2%	3.4%	5.1%	1.2%	0.4%	2.8%
Other	15.9%	5.5%	2.3%	9.4%	6.2%	9.8%	4.6%	7.9%

Table 16 also describes characteristics of household respondents, taken from completed Wave 3 and Wave 4 Household Questionnaires. The interviewer instructions stated that "the household questionnaire should be completed by a household member who is over age 18 and is knowledgeable about all members of the household, including the financial situation of the household". Across all groups in both waves, females are much more likely to be interviewed for the household questionnaire than males. However, employment status is different across the three groups with respondents in white households much more likely to be working. In Wave 3 household heads and their spouses compose more than 80% of respondents in

Coloured and white households, whereas in African households many more young adults were also the respondent to the household questionnaire. In Wave 4 young adults were much less likely to be the respondent in African households than in Wave 3. This is probably partly explained by the inclusion of an older adult sample in Wave 4. In households with both a young adult and an older adult, the interviewer was more likely to interview the older adult for the household questionnaire.

5.4. Wave 4 Older Adult response rates

In Wave 4 the CAPS was expanded to include a panel of older adults. We attempted to interview all original Wave 1 household members who would have been aged 50 or older on 1 January 2006. Response rates by population group and a break-down for the reasons for non-response are shown in Table 17. The response rates for African and coloured older adults were 72% and 74% respectively. As with the younger adults the response rate for white older adults was much lower at 36%. These response rates are calculated as a percentage of all attempted interviews that were successful and include those who had died by the time of the Wave 4 interview. We have chosen to present this simple response rate as we have no idea how many of the non contacts and those classified as moved to an unknown destination had also died by the time of the Wave 4 fieldwork.

Overall the most common reason for non-response is a refusal (38%), followed by moving to an unknown destination (16%), not available (15%) and deceased (15%). Similar to the young adults, refusals are much lower amongst the African older adults. Around a fifth of African and coloured older adults who were not successfully interviewed had died. In contrast deaths were only responsible for less than 5% of the non-response white older adults. The most common reason for non-response for African older adults in moving within South Africa, mainly to the Eastern Cape, and moving to an unknown destination.

The final number of successful interviews were 741 Africans, 1284 coloureds and 252 whites. For every successfully completed older adult interview, the interviewer was instructed to conduct a household questionnaire. Household questionnaires were completed for 99% of the older adult sample.

Table 17: Older Adult response rates

	African	*Coloured*	*White*	*Total*
Attempted interviews	1035	1748	698	3481
Successful interviews	741	1284	252	2277
Response rate	71.6%	73.5%	36.1%	65.4%
Reasons for non-response				
Not Available	4.4%	14.9%	22.1%	15.0%
Refused	18.0%	39.1%	49.9%	38.0%
Deceased	20.8%	20.3%	4.7%	14.6%
Moved within Cape Town	0.3%	0.2%	0.2%	0.3%
Moved within South Africa	25.9%	2.6%	1.4%	7.8%
Moved abroad	0.7%	0.7%	2.0%	1.2%
Moved no details	23.1%	13.4%	13.8%	15.9%
Institutionalised	1.0%	1.1%	0.0%	0.7%
Mentally Unfit/Disabled	2.7%	4.3%	1.8%	3.0%
No contact	3.1%	3.5%	4.1%	3.7%
Completed OAs with completed household Qs	99.6%	98.0%	98.8%	98.6%

Table 18: Characteristics of OA sample

	African	*Coloured*	*White*	*Total*
Sex				
Male	42.7%	39.3%	49.2%	41.5%
Female	57.3%	60.7%	50.8%	58.5%
Age category				
50 to 54	32.7%	30.1%	37.9%	31.8%
55 to 59	26.8%	23.6%	19.8%	24.2%
60 to 69	26.7%	29.8%	23.0%	28.0%
70 to 79	10.6%	13.0%	12.9%	12.2%
80 plus	3.1%	3.6%	6.5%	3.7%
Currently earning income				
Yes	47.1%	37.7%	54.4%	42.6%
No	52.9%	62.3%	45.6%	57.4%

Table 18 presents characteristics of the older adults who were successfully interviewed in Wave 4. African and coloured older adults are more likely to be female while there are roughly equal numbers of men and women in the white sample. Over half the sample is below the age of 60 and over 80% of the sample is below the age of 70. Just over 40% of the sample report that they are currently doing something to earn money. White older adults are the most likely to report working.

Table 19 examines the determinants of non-response in a multivariate context. The table presents marginal effects and standard errors from a probit regression in which the dependent variable is set equal to one if the older adult did not successfully complete an older adult questionnaire in Wave 4, and is set equal to zero otherwise. This non-response variable is regressed on dummy variables for years of age, age squared, gender, population group, years of completed schooling, quintiles of per capita income, indicators that the household has an indoor toilet, water and electricity and an indicator that a young adult lived in the household. All characteristics are taken from the Wave 1 household data. In addition to the income quintiles a dummy is included for missing income. The marginal effects can be interpreted as the increase (decrease) in the probability of non-response from being in that category, controlling for the other characteristics. For example, in African older adults are estimated to have response rates that are 25% higher than white young adults, evaluated at the mean values of the other characteristics.

Table 19: Marginal effects and standard errors from a probit regression for Older Adult non-response in CAPS Wave 4.

Female	-0.087	[0.017]**
African	-0.252	[0.029]**
coloured	-0.291	[0.026]**
Age	-0.035	[0.010]**
Age squared	0	[0.000]**
Years of completed education	0.001	[0.003]
Indoor toilet	-0.055	[0.052]
Electricity	-0.15	[0.049]**
Indoor water source	-0.023	[0.032]
Per capita income quintile 2	-0.028	[0.030]
Per capita income quintile 3	0.013	[0.032]
Per capita income quintile 4	0.009	[0.034]
Per capita income quintile 5	0.061	[0.038]
Missing income	0.098	[0.044]*
Young adult in Wave 1 household	-0.127	[0.020]**
Observations	3257	

Standard errors in brackets

* significant at 5%; ** significant at 1%

Higher non-response is observed among men, whites, younger respondents and people in households with electricity. Interestingly the relationship between per capita income quintile and non-response is not significant. However, significantly higher non-response is observed for household with missing household income. Older adults from households that contained individuals in the young adult sample are 13 percentage points more likely to have a successful Wave 4 interview. This is not surprising as these households would have been revisited in Waves 2 and 3.

5.5. Wave 4 Child response rates

The child sample is defined as all children born to female young adults who were successfully interviewed in Wave 4. A total of 921 children were identified from Wave 4

young adult interviews and 834 or 91% of these children were successfully contacted. Only 2 white children were identified. This is due to a combination of poor initial response, high attrition, lower fertility and higher age at first birth for white females. The response rate was lowest (88%) for African children. In most cases where these children were not interviewed it was because they were not co-resident with the young adult and were living with other family in the Eastern Cape. Of the successfully interviewed children, 97% have a corresponding household interview.

Table 20: Wave 4 child response rates

	African	*Coloured*	*White*	*Total*
Interview attempted	481	438	2	921
Interview Completed	421	411	2	834
Response rate	87.5%	93.8%	100.0%	90.6%
Completed children with completed household Qs	98.6%	95.9%	100.0%	97.2%

There are roughly equal numbers of boys and girls in the sample and the average age is 2.7 years old.

6. Weights

6.1. Wave 1 weights

The public release data sets include sample weights that should be used to adjust for the sample design summarized in Section 2. Three sample weights for Wave 1 are included in the data, each dealing with specific issues.

The first sample weight, *weightsd*, adjusts for three critical elements of the sample design: 1) the intentional oversampling of African and white households; 2) the intentional differential sampling of households with and without young adult household members; and 3) the addition of secondary households (backyard shacks) into the sample of screener households in the field. This weight is incorporated into the other two sample weights. It can be considered as either a household or individual weight if there is no concern with adjustment for non-response.

The second weight, *weighthr*, begins from the first weight and adds additional adjustments for unit non-response at the level of PSUs. In order to avoid unusually large weight being given to households in PSUs with low response rates, a number of PSUs in the same population group cluster which were in close geographic proximity were combined for purpose of adjusting for unit non-response. This should be considered the appropriate household weight if it is considered desirable to adjust for non-response. The implicit assumption in the adjustment for non-response is that the households that responded to the interview do not differ systematically from the households in the same PSU that did not respond. While this assumption is unlikely to strictly be true, it is true that most enumeration areas are relatively homogenous neighbourhoods. Since we have no information on households that did not respond to the interview, there is no way to explicitly examine the extent to which they differ from responding households.

The third sample weight, *weightyr*, is an individual young adult weight that adds additional adjustment for individual non-response. This adjustment is made by calculating response rates for each combination of single years of age, sex, and population group (8x2x3=48 cells) using the information provided on the household questionnaire. The small number of individuals classified as Indian and other were merged with the coloured group. This approach is taken as an alternative to using young adult response rates at the PSU level, based on the assumption that there is more homogeneity for our purposes among all white 18 year-old males in Cape Town than there is among the 14-22 year-olds in a given PSU. As discussed above, the response rates were lower for older white and coloured males, so the non-response adjustment is greatest for those groups. This weight makes the same implicit assumption about household level non-response as the previous weight, and adds the additional assumption that within a given age/population group/sex cell there are no systematic differences between respondents and non-respondents.

Using the third weight, *weightyr*, the weighted distribution of 14-22 year-olds by population group is within one percentage point of the population group distribution in Cape Town in the 1996 census. This weight should therefore provide results that are reasonably representative of the young adult population of Cape Town.

6.2. Weighting for Wave 2, 3 and 4 Young Adult non-response

In addition to the three sample design weights, the Waves 1-2-3-4 public release data sets include additional weights to adjust for individual young adult non-response in Waves 2, 3 and 4.

Since Wave 2 is composed of two sub-waves (Waves 2a & 2b) with different modules asked of different sub-samples, there are three Wave 2 attrition weights. The weight *w2a_weightyr* corresponds to the Wave 2a sub-sample (approximately one-third of the total CAPS Young adult sample), the weight *w2b_weightyr* corresponds to the Wave 2b sub-sample (approximately two-thirds of the total CAPS Young adult sample), and the weight *w2y_weightyr* corresponds to the combined "total" Wave 2 sample. All of these weights are individual young adult weights that add an additional adjustment for individual young adult non-response in Wave 2a, 2b or 2 "total" to the weight *weightyr,* which adjusts for the sample design and Wave 1 non-response.

Similarly the weights, *w3y_weightyr* and *w4y_weightyr,* are individual young adult weights that add additional adjustment for individual young adult non-response in Waves 3 and 4 to the weight *weightyr.*

The adjustment for Wave 2a, 2b, Wave 2 "total", Wave 3 or Wave 4 young adult non-response is made by estimating separate probit models of the probability the respondent completed a Wave 2a, 2b, either of the Wave 2, Wave 3 or Wave 4 young adult questionnaire. Information given in Wave 1 on age, sex and population group was included in the model. As in the construction of the original weight *weightyr*, the small number of individuals classified as Indian and other were merged with the Coloured group. From the estimation, the predicted probability was inverted and then capped at the 99% percentile to obtain the non-response adjustment.

These weights makes the same implicit assumptions about Wave 1 household and young adult level non-response as the weight *weightyr*, and add the additional assumption that within age/population group/gender groups there are no systematic differences between respondents and non-respondents to each of Waves 2a, 2b, 3 and 4.

Common trends[2] in response rates by sex, population group and age are observed in all waves. Response rates were lowest for older white respondents, with no significant difference in response rate between males and females. As such, the non-response adjustment is greatest for these groups of young adults in each wave.

6.3. Weighting for Wave 4 Older Adult non-response

In addition to the household level sample design weights, the Older Adult data set includes a weight, *weightor,* to adjust for individual older adult non-response in Wave 4. This adjustment is made by calculating response rates for each strata defined by single years of

[2] There is one exception; African response rates (of those who responded in Wave 1) are significantly lower than Coloured response rates in Wave 3 and 4 but not in Wave 2.

age and age squared, sex, and population group using the information provided on the household questionnaire updated with Wave 4 information if this information was found to be different[3]. The small number of individuals classified as Indian and other were merged with the coloured group. This approach is taken as an alternative to using older adult response rates at the PSU level, based on the assumption that there is more homogeneity for our purposes among all white 65 year-old males in Cape Town than there is among the >56 year-olds in a given PSU. The adjustment for Wave 4 older adult non-response is made by estimating a probit model of the probability the respondent completed a Wave 4 older adult questionnaire. From the estimation, the predicted probability was inverted and then capped at the 99% percentile to obtain the non-response adjustment. Response rates were lowest for younger[4], white, male older adults. As such, the non-response adjustment is greatest for these groups of older adults.

In the construction of this weight, two implicit assumptions are made. First, households that responded to the interview do not differ systematically from the households in the same PSU that did not respond. Second, within a given age/population group/sex cell there are no systematic differences between respondents and non-respondents. Use of this weight in the *capsw4.o.v.dta* should provide results that are reasonably representative of the older adult population of Cape Town in 2002.

6.4. Weighting in general

Users of the CAPS data should use one of these sample weights in order to adjust for the key features of the sample design. The weights all adjust for the systematic oversampling of African and white households, and for the differential sampling probabilities for households with and without young adults.

The household, young adult and older adult weights which account for non-response make particular assumptions about non-response, which users may or may not want to assume.

[3] In the Wave 1 household questionnaire the Older Adult might not have been the household respondent. The assumption is made that demographic information from the Wave 4 Older adult questionnaire is more reliable than information from the Wave 1 household questionnaire.

[4] Response rate increase at a decreasing rate with age

7. Questionnaires and content

7.1. Wave 1

7.1.1. Questionnaire design

The CAPS Wave 1 questionnaire was developed in 2001 and 2002 through a series of meetings between UCT and UM collaborators held in Cape Town and Ann Arbor, Michigan. Questionnaires covering similar themes from studies in South Africa, the United States, and other countries were consulted, with consideration given to issues of comparability and consistency with these other studies whenever possible. The questionnaire was tested in small groups and in two pilot surveys held in May and June 2002.

Three questionnaires were administered in CAPS Wave 1 – a household questionnaire, a young adult (YA) questionnaire, and a young adult literacy/numeracy evaluation (LNE). For the household and young adult questionnaires, the interviewers filled out an English version of the printed questionnaire, but had access to Xhosa and Afrikaans translations of the questionnaires for use in asking questions. Respondents could choose English, Afrikaans, or Xhosa as the language used during the interviewed. The language in which the interview was conducted is recorded in the data. The literacy and numeracy evaluation was completed by the young adult respondent, who chose either an English of Afrikaans version of the evaluation. The language of the LNE is recorded in the data.

The household questionnaires collected data on all members of the household, covering basic social and demographic variables, education, migration to Cape Town, work and income. Membership of the household was defined in terms of 'usually' living in the household, meaning that 'the person has lived here for more than 15 days of the last 30 days'. This is a less restrictive definition than the one used by Statistics South Africa (in its October Household Surveys, for example) where a household comprises the people who 'eat together and share resources' and 'normally resides at least four nights a week' at the place of interview. We chose to relax the 'common pot' and 'pooled resources' criteria because we, and many others, worry that, in some cases, people who live together (in terms of sleeping under the same roof) might not eat together and might not share resources. The assumption that co-residence implies these other things is derived primarily from the western experience of nuclear family households, and is invalid in many Southern African cases (see Russell, 2003a, 2003b).

At the same time, we chose not to use the more elastic time stipulation used, for example, in the 1993 PSLSD where a household comprises people who live together, under the same roof or in the same homestead or compound, for 'at least fifteen days out of *the past year*' (the PSLSD also employed the common pot and shared resources criteria). The PSLSD time stipulation made sense for a countrywide sample that wanted to collect data on migrant workers in their households of origin. In Cape Town, which is a destination not a source of migrant workers, the looser criterion was not necessary.

The young adult questionnaire collected a mix of current and retrospective data on the lives of our respondents. Much of the retrospective data were collected in the form of a life-

history calendar, which recorded year-by-year details of schooling, who the respondent had lived with, pregnancies and births, from birth to the present.

7.1.2. Wave 1 Household Questionnaire

Once a household was selected into the CAPS sample (following the rules described above in the section on sample design), one person was selected to complete the household questionnaire. Interviewers were instructed that "the person answering the household questionnaire should have the most knowledge about everyone in the household and must be over the age of 18."

Roster of household members

The household questionnaire begins with a complete roster of all members of the household. The questionnaire asked about individuals "who usually live here," which was defined as having spent more than 15 days out of the last 30 days "living under this roof." The roster includes information on variables such as age, schooling, migration history, work status, and income (for more details see Table 22). The household roster included questions about the occupation of each household member. These answers sometimes included specific names of businesses which if released in the data might compromise the confidentiality of respondents. For this reason this information has been replaced with standard international occupation codes using the Standard Occupational Classification System (SOC) and Standard Industry Classification System (SIC), with modifications for use in South Africa.

The household roster was used as the basis for selecting young adults into the young adult sample. Up to three household members age 14-22 are identified in the roster, and questions are asked about the relationship of all other household members to each young adult. In cases where more than three individuals age 14-22 lived in the households, interviewers were instructed to select the three with the most recent birthdays for the young adult sample.

Roster of non-resident children

After completion of the household roster, a roster was collected of all biological children aged 0-22 of household members who were not living in the household. Information was collected on variables such as age, schooling, work activity, and location of current residence of these non-resident children, and the line numbers of the corresponding mother and/or father from the household roster were recorded.

Household Events

Module B of the household questionnaire collected information on events that had affected the household in the previous 24 months. These included the death or serious illness of household members, loss of employment or financial support, and abandonment or divorce.

Household Characteristics

Module C of the household questionnaire collected information on household characteristics such as the type of dwelling, access to water and electricity, and ownership of consumer durables.

Household Income, Expenditure and Debt

Module D collected information on household income, expenditure, and debt.

7.1.3. Wave 1 Young Adult Questionnaire

Each of up to three household members age 14-22 was asked to complete the young adult questionnaire. As noted above, in cases where more than three young adults lived in the household, the three with the most recent birthdays were selected to complete the young adult questionnaire. Written parental consent was obtained for all individuals under the age of 18, in addition to the written consent of the young adult respondent.

Some of the information in the young adult questionnaire overlaps with information in the household roster. Both questionnaires, for example, include information on the age, population group, school attendance, and work activity of the young adult. Interviewers and supervisors were told not to consider answers incorrect simply because they were inconsistent between the two questionnaires. In many cases inconsistencies did prompt follow-up contact our double-checking of questionnaires, especially when there were large discrepancies in variables such as age. There continue to be a number of cases in which answers in the young adult questionnaires do not agree with information provided for that young adult in the household questionnaire. For some variables, such as school or work activity, this may be because the household respondent was not fully aware of the young adult's activities. In other cases, such as population group, the young adult may self-identify in a way that is different than the population group given by the household respondent. In most cases it is impossible to know the source of the inconsistency, or to say with certainty which answer is correct. We assume that answers provided directly by the young adult are more likely to be correct, but researchers can make their own decisions about how to deal with the relatively small number of inconsistencies between the two questionnaires.

The young adult (YA) questionnaire contained the following modules:

Background Characteristics

Module A of the YA questionnaire collected information on basic demographic characteristics, including age, population group, region of birth, primary language, and religion.

Life History Calendar

One of the most important features of the YA questionnaire was the life history calendar. This calendar collects information related to many of the specific modules in the questionnaire. This calendar is discussed in detail below.

Schooling

Module C collected information on current and previous schooling activity, supplementing the information on schooling collected in the calendar. This module included information on the name and location of the current and previous schools attended by the respondent. In order to protect the confidentially of respondents, the information on names of schools is not included in the standard public release data set. These variables may be made available on a

case-by-case basis to researchers who need the data on school names for specific research projects. Researchers who are interested in working with the data on specific schools should contact the principal investigators.

Employment

Module D collected information on the respondent's labour force activity. This includes information on job search and specific information on up to three jobs – the current or most recent job, the previous job, and the first job. The questionnaire included questions about the names of employers. This was used to help identify the occupation and type of economic activity of respondents and has been removed from the public data sets to protect confidentiality. This information has been replaced with standardized occupation and industry codes.

Because this module collects information on multiple jobs and spells of job search, the skip pattern for this section of the questionnaire is more complicated than it is for other modules. While we have corrected many skip violations using a combination of follow-up with respondents, logical consistency checks, and inspection of the paper questionnaire, we have not attempted to reconcile all violations of skip patterns in the data. In many cases we have created new variables that provide information in a simpler form than the questions from the original questionnaire.

Health and Fertility:

Module E collected information on health conditions, sexual activity, and childbearing, and included a battery of questions related to HIV/AIDS. All of the information was collected through the written questionnaire. No measurements or physical examinations were done and no biological material was collected. First names of children born were collected to simplify the interviewing process. These names have been removed from the public data set. This module includes a number of sensitive questions about sexual activity. Interviewers were instructed that every attempt should be made to administer this section of the interview in private, including going outside if necessary. Interviewers were also instructed that some sections of the questionnaire could be completed directly by the respondent if the respondent seemed uncomfortable answering the questions out loud or if other people were present.

Some sections of this module are completed directly on the calendar. Other questions refer to the timing of events, such as the age of first sexual activity, but do not put that information on the calendar. Interviewers were instructed to check for the consistency in the timing of events across questions, including checks against the calendar. Some inconsistencies remain in the data even after cleaning, however, including some women who report ages of birth that are prior to the age at first sexual activity. These data have been left as is, since it is not clear which of the answers is correct. Researchers interested in the specific timing of events should also carefully read the section describing the calendar and its two different time frames.

Non-Resident Biological Parents:

Module F collected information on the biological parents of the respondent. Since the household roster contains information on parents who are co-resident in the household, the focus of the section was on parents who do not live in the household. Information covered is

similar to that in the household roster, including age, schooling, and employment. It includes questions on the occupation of each living parent (which have been replaced with standardized codes),

Grandparents:

Module G collected information on all four of the respondent's biological grandparents. Information includes age, receipt of government pension, and current location.

Parental Investment:

Module H collected information on the extent of involvement of biological parents, stepparents, and other guardians in the lives of the young adult respondents. Information is collected on variables such as frequency of contact, financial contributions, and involvement in personal matters.

Childhood and Family Environment:

Module I collected information about the environment in which the YA grew up. It includes questions on the influence of individuals on the YA and features of the home environment such as the presence of drugs or violence in the home.

Time Allocation and Social Involvement

Module J included questions on hours spent in the past week doing activities such as paid work, housework, attending school, and caring for children or adults. It also includes questions on involvement with groups such as sports teams, religious groups, or music groups, as well as information on use of alcohol and drugs.

Contact Information:

Module K collected information on people who could be contacted to help locate the respondent for future waves of the panel. This information is confidential and does not appear in public releases of the data.

Interviewer Evaluation

The interview evaluation (Module M) includes details of the interview such as end time and language used and also asks the interviewer to describe the respondent's vocabulary, attitude, and attentiveness as well as the privacy of the interview.

7.1.3.1. Wave 1 Young Adult Life History Calendar

The life history calendar is a major focus of the young adult questionnaire. The structure of the calendar was designed with the goal of capturing information in a way that was natural for respondents and that also minimized errors during the interview or data capture. The original structure of the completed questionnaires is not very useful for analysis by researchers, however, so the data has been transformed in a number of ways to facilitate

research. It is important that researchers understand the design of the calendar and the ways in which the data have been transformed.

The calendar is divided into two sections, each taking up one page of the calendar. These two sections have different time perspectives, a reflection of the different substantive focus of each section. The first section covers household living arrangements and relationships. Each column of the calendar corresponds to a year of the respondent's life, beginning at age 0 and going as high as age 22. Rows of the calendar correspond to different questions in the questionnaire. Information is marked on the calendar in different ways for each question, with many questions having a box to indicate "always" or "never" at the beginning of the relevant row. For example, question B2a asks respondents whether they lived with their biological mother at every age from birth to their current age. A box at the beginning of the row is checked to indicate that the respondent lived with his/her mother always, sometimes, or never. If "sometimes" is chosen, periods of co residence are to be marked by putting an "X" under the age at which co residence began, an "O" in the year that co residence ended, and a line drawn connecting the "X" and "O".

The second page of the calendar focuses on school and work. Since we wanted detailed information on the outcome of each school year, we wanted to focus this part of the calendar on school years. Since the South African school year roughly coincides with a calendar year, it was natural to organize this section based on calendar years. Interviewers were told to write "2002" under the age the respondent was on 1 January 2002 and to fill in other columns by working backwards from 2002. The column of the table headed "12", for example, describes the calendar/school year in which the respondent was age 12 on 1 January. Questions are asked about variables such as whether the YA attended school, whether the YA passed or failed that grade, whether the YA worked that year, and whether the YA became pregnant that year.

We think the choice of time frames in the two sections of the calendar provided the best frame of reference for the young adult respondents, while keeping the interview as simple as possible for both interviewers and respondents. Unfortunately it means that researchers must be cautious in looking at the timing of events. The column marked "12" on the first page of the calendar covers the period in which the YA was age 12. The column marked "12" in the second page of the calendar covers the year in which the YA was age 12 at the beginning of the year. For a person born late in the year these will come close to covering the same period of time. For a person born early in the year, however, the time periods can differ by almost a year. Users should be cautious about this difference, and should use the date of birth of the YA as an additional piece of information when considering the timing of events.

7.1.4. Wave 1 Young Adult Literacy and Numeracy Evaluation

Each young adult respondent was asked to complete a literacy and numeracy evaluation (LNE). This evaluation was developed in consultation with the Joint Education Trust, especially Nick Taylor and Penny Vinjevold. This evaluation took about 20 minutes for most respondents to complete.

The LNE instrument was available in English and Afrikaans. Although speakers of Xhosa or other languages could choose whether to take the English or Afrikaans, over 99% of those who said Xhosa was their main language completed the English version of the LNE.

Comparing results of the evaluation across population groups must therefore be done with caution. For Xhosa speakers the LNE is a test of English language ability in addition to basic literacy and numeracy.

7.2. Wave 2, 3 and 4

7.2.1. Questionnaires

Table 21 presents a list of questionnaires administered in each wave. The young adult questionnaire was administered to all successfully re-contacted young adults from Wave 1 in each of Wave 2 (either 2a or 2b), Wave 3 and Wave 4. In addition, each wave included either a separate household questionnaire (Wave 3 and 4) or information about the household in the young adult questionnaire (Wave 2a and 2b). The household questionnaire was administered at each completed young adult's household. In Waves 3 and 4 some additional questionnaires were included.

Table 21: CAPS Questionnaires by Wave

	Wave 1 (2002)	Wave 2a (2003)	Wave 2b (2004)	Wave 3 (2005)	Wave 4 (2006)
Young Adult Questionnaire	x	x	x	x	x
Young Adult Proxy Questionnaire					x
Young Adult Numeracy & Literacy Evaluation	x				
Household Questionnaire (separate from YA questionnaire)	x			x	x
Parent Questionnaire				x	
Older Adult Questionnaire					x
Child Questionnaire					x

7.2.1.1. Wave 3 Parent Questionnaire

New to Wave 3 was a parent questionnaire, administered to co-resident parents or guardians of young adults. This questionnaire focuses on the parent or guardian's attitudes and beliefs on education and value socialization, their assessment of the home environment in which the respondents had grown up, and their social and economic expectations for and of their children.

The parent questionnaire begins by asking the parent-respondent questions concerning their own happiness and perceptions of control over their own life and future opportunities. The questionnaire then asks the parent about the importance of education, in general. In the next section, the parent respondent is asked questions specifically about the young adults as well as an addition "selected child" from their household who was aged 7-16 at the time of the interview. The questions in this section ask about the expectations the parent/guardian has for the adolescents, their perception of the adolescents' attitudes towards school and work, and the relationships that the parent/guardian has with each adolescent covered in the questionnaire. The questionnaire concludes with questions on the parent/guardian's main

sources of influence and encouragement, life expectancy of the respondent, for both him/herself and all of his/her own living biological children and personal acquaintance with individuals with HIV/AIDS.

7.2.1.2 .Wave 4 Young Adult proxy questionnaire

New to Wave 4 was a young adult proxy questionnaire. In the case where the young adult respondent could not be interviewed an attempt was made to administer a young adult proxy questionnaire to an adult who was knowledgeable about the young adult. The questionnaire collected basic information about the Young adults' whereabouts, education, marital status, health, employment and number of children.

7.2.1.3 .Wave 4 Older Adult questionnaire

In Wave 4 an older adult questionnaire was administered to all residents of original Wave 1 CAPS households who were age 50 or over on 1 January 2006. Table 22 lists the contents of each section in the questionnaire.

Table 22: Content of the Older Adult Wave 4 questionnaire

Section	***Content***
A	Personal information
B	Employment, income, education and Martial Status
C	Roster of children
D	Income and family support
E	Connections to the Eastern Cape
F	Health and health seeking behaviour
G	Habits
H	Functional Status
I	Cognitive Function
J	Measurements
K	Mobility
L	Interview Evaluation

7.2.1.4 .Wave 4 Child questionnaire

The child questionnaire was administered to each biological child of all female CAPS young adults that were interviewed in Wave 4. The questionnaire collects current physical measurements as well as information from the child's road to health card.

7.2.2. Content

7.2.2.1. YA Content

The Young adult questionnaire in each Wave included questions that 1) updated data information collected in previous interviews and 2) added new modules. In waves 3 and 4, information on schooling, work and job search was updated from the date of the last successful interview. Therefore all respondents with both Wave 1 and Wave 4 interviews

completed have uninterrupted schooling, work, job search data from 2002 through to 2006, even those not interviewed successfully in Wave 2 and/or Wave 3.

The young adult questionnaires each included a number of modules. Table 23 presents a breakdown of the modules included in the young adult questionnaire for each wave. Important details of the modules are summarized below.

a. Primary Focus of each wave:

Wave 2

Wave 2a focused on HIV/AIDS, sexual behaviour and attitudes. Wave 2b focused on schooling, including school choice, and economic activity.

In Wave 2a (2003), more comprehensive data on HIV/AIDS was collected, including attitudes around HIV and HIV prevention and risk behaviour. In addition, the job table format was introduced to collect information on all successive jobs since last interview. Note there is no Module D in the Wave 2a questionnaire and no Module J in the Wave 2b questionnaire.

In Wave 2, the household questionnaire was short and included as part of the Young adult questionnaire. A complete roster of all members from the Wave 1 household questionnaire was pre-printed in the Wave 2 questionnaire. For pre-printed individuals no longer resident, the reason why they moved out was asked. Respondents were also asked to update the roster with any current members of the household who were not listed on the pre-printed roster. For each individual, the roster updates information on variables such as age, schooling and work status.

Wave 2a included a module, Module B, for respondents who had moved from their recorded Wave 1 residence. This module collected information on characteristics of the new household such as the type of dwelling, access to water and electricity as well as information regarding the reasons for the move.

Wave 2b included a section, Module D, on events affecting the household and respondent. This module collected information on the health of the young adult as well as on events that had affected the household since August 2002. These included the death or serious illness of household members, loss of employment and start of a new job or grant. Respondents were also asked about deaths of family members who do not live in their household.

Wave 3

Wave 3 covered the same areas Wave 1 had focused on. New modules include a detailed residential and schooling history and a roster of sexual partners for young adults.

The Wave 3 household questionnaire contained much of the same material as in Wave 1, with the addition of modules on household expenditure, illness, death and both outgoing and incoming income transfers.

Table 23: Content of CAPS Young adult questionnaires by Wave

	Wave 1 (2002)	Wave 2a (2003)	Wave 2b (2004)	Wave 3 (2005)	Wave 4 (2006)
Main themes in CAPS	Section				
Demographics/ Personal Information	A	A	A	A, B	A
Location/Migration		B	B	B	B
Schooling	C, B	E	E	B, C	C
Job Table		F	F	D	D
Employment-salaried	D, B	F	F	D	D
Employment-self-employed			G		
Job search	D	F	H	D	D
Unpaid/domestic work			I		
Health	E			E	E
Fertility	E			E	F
Marriage	E, B			E	E
Sexual relationships and HIV/AIDS	E	H		E	E, H
Attitudes on sex and marriage		G			
Attitudes on HIV/AIDS		C, J			J
Attitudes to others		I			
Non-Resident Biological Parents	F				
Grandparents	G				
Parental Investment	H				
Childhood and Family Environment	I, B				
Relationships with parents/other adults				F	
Family support and kin					G
Time Allocation and Social Involvement	J			G	
Physical Measurements					I
Interview Evaluation	M	L	K	H	K
Life History Calendar	B				
Month-by-month calendar:		X	X	X	X
Moving residence			B3-B5	B16-B18	
Health and disability			D4	E6, E8	E31
Household shocks			D5-D10		
Schooling		F1	E1	C4	C10
Change schools				C8	C20
Work		F2, F3, F4, F5	F9	D8	D8
Job search		F6, F7	F10, F11	D26	D16
Marital events					E47
Residential and Schooling History				B	
Household roster		C	C		
Events affecting the household and respondent			D		

Wave 4

In addition to providing follow-up information on the school, work, and childbearing histories of CAPS young adults, Wave 4 expands the focus on health and systems of family support. The Wave 4 household questionnaire is very similar to the Wave 3 household questionnaire.

b. Details of some modules first included post wave 1:

Two question formats were first introduced in Wave 2 and are common to waves 2, 3 and 4. First, Wave 2 included an expanded monthly calendar. Second, the job table format for capturing labour force activity was introduced.

Monthly Calendar

The monthly calendar is part of young adult questionnaire and captures monthly information on moving residence, attending/changing school, work/looking for work, illness that interfers with normal activities, deaths and in wave 4 marital status. See Table 6.3a for details of which information was included in each wave. The calendar begins from August 2002 and continues through to the month of the respondent interview. From Wave 3 onwards the calendar was updated from the month of the respondents' last interview.

Job Table

The job table in Wave 2 required the respondent to list all jobs held since August 2002. From Wave 3 onwards respondents were required to list information about all jobs since last interview including any jobs the respondent was still working at in the last interview.

Questions included the names of employers, wages, how the respondent got the job and if they had stopping doing this work, the reason why they had stopped, were asked for each job. Names of employers and place of work is replaced with standardized occupation and industry codes, which are available on the CAPS website.

Residential and Schooling History

Module B in Wave 3 is a residential and schooling history.This module included information on all places lived since age 14, the name and location of all schools attended by the respondent, and grade level and school change information in each year of schooling. Questions regarding residential moves are recorded on the Monthly Calendar. In order to protect the confidentially of respondents, the information on names of schools and towns is not included in the standard public release data set. These variables may be made available on a case-by-case basis to researchers who need the data on school names for specific research projects. Researchers who are interested in working with the data on specific schools should contact the principal investigators.

Family support and kin

Physical Measurements

7.2.2.2. HH Content

The household questionnaires each included a number of modules. Table 24 presents a breakdown of the modules included in the household questionnaire for each wave.

Table 24: Content of CAPS Household questionnaires by Wave

	Wave 1 (2002)	Wave 2a (2003) †	Wave 2b (2004) †	Wave 3 (2005)	Wave 4 (2006)
Roster of household members:					
Demographics/personal information	A	C	C	B	B
marital status	A			B	B
schooling	A			B	B
migration	A			B	B
work and income	A			B	B
grants	A			B	B
job search	A			B	B
parental residence	A			B	B
health				B	B
Roster of non-resident children	A				
Household events	B		D	D	D
Household characteristics	C	B*		C	C
Household Income, Expenditure and Debt	D			D	D
Income transfers				F	F
Interview Evaluation	F			G	G

†Household questionnaire part of young adult questionnaire
*only for new residences

In Wave 3 and 4 a separate household questionnaire was administered to an adult household member at the current residence of the Young Adult, and included a household roster, and modules on household income, expenditures and transfers in and out of the household. In some cases the young adults in our panel will be appropriate sources of this household-level data: some of our young adults are themselves heads of households or breadwinners. In other cases, parents or other older household members were interviewed. While much of the Wave 3 and 4 household questionnaire is identical to Wave 1, there are a few new additions. These new household-level modules cover: expenditure; events and shocks affecting the household, such as death, serious illness, loss of job, crime, etc; income transfers in and out of the household. The roster includes new questions on health, grants received and school fees paid by household members. See Table 24 for details

7.3. Pre-Loaded and Pre-edited Information

Given the nature of the panel, there were some questions which did not need to be asked again of all of the respondents in each wave. Furthermore, respondents may have been

interviewed at different points in the panel due to design[5]or non-response. As a result, the last date of previously recorded data varied across respondents when interviewers went into the field in waves 3 and 4. This led to a different starting point for the Monthly Calendar in Wave 3 and 4. Sections such as schooling and the household roster required pre-loading of information in order to retain consistency in the panel and also to establish a starting point. Finally, some modules had different versions of the module's questions by population group. For all of these reasons, there are many questions within each of the questionnaires for which we have "pre-loaded" an answer or a starting point in time, shaded out an unnecessary field, or deleted some questions completely.

The questionnaires which are available for download from the website have an example of these "pre-loaded" fields which appear in italic or bold type.

Table 25 presents all fields/variables that were preloaded in the young adult questionnaire. The table is divided into three sections. Section 1 indicates fields/variables where information detail from a previous wave was preloaded into the questionnaire before the interview went into field. Section 2 indicates variables where the question was asked since the last interview or variables specific to dates which did not apply to the respondent or had already been asked were shaded out on the questionnaire. The final section, Section 3, indicates variables/sections which were asked only of certain groups or were deleted from the questionnaire if the information had previously been collected.

Table 26 presents all variables that were preloaded in the household questionnaire. There are many pre-loaded fields for each household member who was resident in the household (defined as living in the household for more than 15 out of the last 30 days) at the time of the last successful interview. Pre-loaded information in the household roster is intended to facilitate the matching of household members across waves.

[5] While all respondents were interviewed in Wave 1 (2002), about 1000 were successfully re-interviewed in Wave 2a (2003) and about 2600 were successfully re-interviewed in Wave 2b (2004).

Table 25: Preloads-Young adult questionnaire for Waves 2a, 2b, 3 and 4

	Wave 2a (2003)	Wave 2b (2004)	Wave 3 (2005)	Wave 4 (2006)
Preloads and Pre-edits				
1. Information				
Personal Information				
Name*	x	x	x	x
Community*	x	x	x	x
Sex		x	x	x
Population group		x	x	x
Address*	x	x	x	x
Work telephone*	x	x	x	x
Home telephone*	x	x	x	x
Cell phone*	x	x	x	x
Email*	x	x	x	x
Nicknames*	x	x	x	x
Prefered language			x	x
Information of three contacts*			x	x
Interview information				
Person ID	x	x	x	x
Wave 1 household ID	x	x	x	x
HHID extension			x	
Original enumeration area*	x	x	x	x
Still at Wave 1 adress				x
Wave 1 area*				x
Wave 1 community*				x
Year of last interview			x	x
Whether the YA was interviewed in 2A or 2B			x	x
Schooling				
Whether attending school	E.1a	A.13		
School Name	A.11/E.1a	A.13	B.9	
Grade	E.1b	A.14	B.12	
Whether attending post secondary schooling	E.2	A.15		
Not enrolled	E.3	A.16		
Residential and schooling history				
Age in residential and schooling history- starting with zero in year of birth.			B.6	
Place of residence-for yeasr when last interviewed			B.7	
Job Table				
Kind of work			D.3	D.3
Name of employer			D.4	D.4
Main business of place of work			D.5	D.5
Month started doing this job				D.7m
Year started doing this job				D.7y
How respondent got job			D.10	D.10
Children				
Age at pregnancy			E.35	F.2
Age of person made you/you made pregnant				F.3
Outcome of pregnancy			E.38	F.6
Name of child			E.39	F.7
Gender of child			E.40	F.8
Date of birth of child			E.41	F.9
Family Support and kin				
Relationship of parents, OAs and other YAs to respondent				G.6.
Did … help you with domestic chores etc				G.13.2a
Did you help … with basic personal activities				G.14.2a
Does … send you money or goods in a usual month				G.15.2a
Did … contribute to pay for large expenses				G.16.2a
Did you send money or goods to … in the past 12 months				G.17.2a

		Wave 2a (2003)	Wave 2b (2004)	Wave 3 (2005)	Wave 4 (2006)
Preloads and Pre-edits					
2. Starting Position-shaded\last interview date					
Residential and schooling history					
	Place of residence: shaded through age 13			B.7	
	School name: shaded through age 4			B.9	
	Name of suburb where school located: shaded through age 4			B.10	
	Grade/level: shaded through age 4			B.12	
	Reason for changing institution: shaded through age 4			B.13	
Schooling and Higher Education					
	Attended school/classes of any kind since last interview			C.3	C1/C19
	Received results for any schooling since last interview			C.3a	
	Record school attendance on monthly calendar, since last interview			C.4	C.20
	Schooling at primary or secondary level since last interview			C.4a	
	Enrolled in higher education that requires matric since last interview			C.34	
	Enrolled in higher education that does not requires matric since last interview			C.34a	
	Enrolled in grade 12 since last interview				C14
	Written matric exam or received results since last interview			C.13	C15
	Year wrote matric, including year of last interview			C.14	
	Month started new school since last interview			C.8	C.10
Schooling & higher education Table(s)					
	Beginning since date of last interview			C.5-C.12/C.35-C.40	C.1-C.12
Employment and job search					
	Done any work since last interview			D.2	D2a
	Working at last interview			D.2	D2b
	Months working since last interview			D.8	D.8
	Looked for work since last interview			D.25	D15
	Months looking for work since last interview			D.26	D16
Children					
	Age of person made you/you made pregnant (shaded)			E.36	
	Timing of pregnancy (shaded)			E.42	
Other					
	Moved since last interview			B.14	B1
	Months when moved residence since last interview			B.16/17/18	
	Illness or serious injury since month of last interview			E.5	
	Months ill health or disability interfered with normal activity since last interview			E.6	E.31
	Death of family member outside the household since last interview			E.7	G.18
	Month of death of family member outside the household since last interview			E.8	
	Marital events since last interview				E.47

*Not included in public release data

	Wave 2a (2003)	Wave 2b (2004)	Wave 3 (2005)	Wave 4 (2006)
Preloads and Pre-edits				
3. Inclusion/exclusion of modules				
Photographs	C.25. Photographs A and B were inserted into questionnaires for African/coloured young adults. Photographs C and D were inserted into questionnaires for white young adults			
School choice*		E.62. African and coloured young adults were asked majority population group of area of their school. White young adults were asked if school government or private. E.63. and E.64 Respondents asked why they attend school in area as given in E.62.		
Puberty			E.9a asked ONLY of girls, E.9b asked only of boys. Questions E.9a and E.9b omitted from the questionnaire if answers have been recorded in previous interviews	
First Sex			Questions E.10-E.15 omitted from the questionnaire if answers have been recorded in previous interviews	
Attitudes to HIV/AIDS				Module J: questions ONLY asked for respondents interviewed in 2A

Table 26: Preloads-Household questionnaire for Waves 2a, 2b, 3 and 4

	Wave 2a (2003)	Wave 2b (2004)	Wave 3 (2005)	Wave 4 (2006)
Line number	C.1	C.1	B.1	B.1
Name	C.2	C.2	B.3	B.3
Young Adult Number			B.4	B.4
Older Adult Number				B.4
Age (2002)	C.4	C.4		
Year of birth	C.5	C.5	B.7	B.7y
Sex	C.6	C.6	B.9	B.9
Population group			B.16	B.20
Line number of biological mother if co-resident at last interview			B.10	B.10
Line number of biological father if co-resident at last interview			B.11	B.11
Head of household	C.3	C.3		
Relationship to head of household			B.12	B.12
Relationship to and Name of YA1	C.7	C.7	B.13	B.13
Relationship to and Name of YA2			B.14	B.14
Relationship to and Name of YA3			B.15	B.15
Relationship to and Name of OA1				B.16
Relationship to and Name of OA2				B.17
Relationship to and Name of OA3				B.18
Relationship to and Name of OA4				B.19
Marital status			B.17	B.21
Line number of spouse if applicable			B.18	B.22
Place of birth			B.19	B.23
Year when household member moved to Cape Town			B.20	B.24
Last year studying/working, etc	C.10	C.10		

8. Keeping track of individuals and households

This section provides background information as to how individuals and households are linked over the panel. This was accomplished through the use of public identifiers and by preloading the questionnaires, with special consideration for the splitting of households.

8.1. Public Identifiers

The CAPS public release data sets are presented at the level of individuals (young adults or household members) except in the long versions of the calendars, when a record in the data corresponds to a month or year in the life of a young adult respondent. Included in all datasets are public identifiers to identify individuals and households, but which contain no information by which to identify the household or individual outside of the CAPS data.

In Wave 1, all households were assigned a four-digit neighbourhood identifier (*cluster*), and a four-digit household number within the cluster (*hhnum*) by the sampling team. These *clusters* represent the enumeration areas from the 1996 South Africa Census, but the *cluster* codes appearing in the public data do not match the census enumeration area codes and cannot be used to match CAPS to the census.. The household identifier (*hhid*) is an eight-digit number that uniquely identifies all households in the Wave 1 data; the first four digits represent the *cluster* and the second four digits represent the *hhnum*.

All household members present in Wave 1 were also assigned an eleven-digit unique person identifier (*personid*) in the Wave 1 household questionnaire, which combines each individual's *hhid* and line number from the household roster (*pcode*). The first eight digits of *personid* are the individual's *hhid*, the next two digits are their *pcode* and the final digit is an additional zero.

For non-resident biological children in the capsw1.h.nrc.*v*.dta file, *personid* is again the unique identifier. The line number from the roster of non-resident biological children in the Wave 1 household questionnaire is *pcodenr* for these individuals. The difference in the construction of the *personid* for these individuals is that an extra zero is added after the *hhid*, and before the two-digit *pcodenr*, which ranges from 1-10. In the cases where *pcodenr* is equal to ten, it is set to zero to construct the *personid*, so that it remains unique across all files.

In all Waves 1-2-3-4 data files, all Young Adults, Older Adults, Children of female young adults and other household members continue to be uniquely identified by *personid.* However, the household identifier *hhid* no longer uniquely identifies households in the Waves 1-2-3-4 data. The original *hhid* remains unchanged, based on the original sample design, and stays with a respondent throughout the panel, even as households split in the years following 2002. In the Waves 1-2-3-4 data, splits from an original Wave 1 household (and *hhid*) can be identified as unique households through a combination of the original *hhid* and new household extension (*w2h_hhext*, *w3h_hhext* and *w4h_hhext*) variables. Further discussion of household identifiers across the panel follows in the next section.

8.2. Households across waves

8.2.1. New household formation

As discussed in the section above, the household identifier *hhid* always refers to the original Wave 1 household, and is therefore no longer unique if households split in subsequent waves. These split households can be uniquely identified in the Waves 1-2-3-4 data using the new household extension (*w2h_hhext* and *w3h_hhext*) variables in combination with the original *hhid*.

The household extension variables (*w2h_hhext* for Wave 2, *w3h_hhext* for Wave 3 and *w4h_hhext* for Wave 4) are two-digit numbers which distinguish splits within an original Wave 1 household. The variable *w2h_hhext* has a value of 21, 22 or 23. The first digit (2) identifies the household split in Wave 2. The last digit (1, 2 or 3) is a counter. Since up to three YA's were originally interviewed in a household, the counter is based on the lowest number given to the young adults (*yanum*) resident in the household. For waves 3 and 4, *w3h_hhext* and *w4h_hhext* are constructed similarly with "3" and "4" as the first digit.

Combining the original *hhid* along with zero and then the wave-specific household extension (*w2h_hhext w3h_hhext w4h_hhext*), a unique eleven-digit household identifier has been generated for Waves 2, 3 and 4 (*w2h_hhid w3h_hhid w4h_hhid*)**.**

8.2.3. New household members

In addition to households splitting, there are many new members present in Wave 2, 3 and 4 households that were not present in the previous wave(s). During fieldwork, these new household members are given a line number on the household roster beginning with 31, 32, 33, etc for Wave 2a (2003), 41, 42, 43, etc for Wave 2b (2004), 51, 52, 53, etc for Wave 3 (2005) and 61, 62, 63, etc for Wave 4 (2006). Because of new household formation, there is the possibility of new members in two split-off households both receiving the same line number in the field. In these cases, a third digit is added to the line number to create the *pcode* therefore making it unique within the original household (*hhid*). This third digit is a 1, 2 or 3 according to lowest *yanum* of resident young adults in the current household[6].

Table 27 presents an example of both household splits and new household members. This household of four members, of which three are young adults, has split into two households when the Wave 2a interview is conducted. Additionally, each of these split households includes one new household member who was not present in the Wave 1 household. Following the procedure explained above, the Wave 2 household extension (*w2h_hhext*) assigned to each household is based on the lowest young adult number present in the household and the relevant wave. For "split a", this results in a *w2h_hhext* of "21" and for "split b", this results in a *w2h_hhext* of "23". Subsequently, this second digit is also added to the *pcode* of the new household members in each household. The combination of *hhid* and *pcode*, used to create *personid*, remains unique for these household members and the combination of *hhid* and *w2h_hhext* now uniquely identifies both households in the Wave 2 data.

[6] The variable *yanum* gives the number (1, 2, or 3) of the YA in the household, including those who did not complete questionnaires.

Table 27. Household splits and new household members

Wave 1:					
personid	**hhid**	**pcode**	**yanum**		
12345678010	12345678	1	1		
12345678020	12345678	2			
12345678030	12345678	3	2		
12345678040	12345678	4	3		
Wave 2(a): Household splits into two					
Split a					
personid	**hhid**	**pcode**	**yanum**	**w2h_hhext**	**w2h_hhid**
12345678010	12345678	1	1	21	1234567821
12345678020	12345678	2		21	1234567821
12345678030	12345678	3	2	21	1234567821
12345678311	12345678	311		21	1234567821
Split b					
personid	**hhid**	**pcode**	**yanum**	**w2h_hhext**	**w2h_hhid**
12345678040	12345678	4	3	23	1234567823
12345678313	12345678	313		23	1234567823
personid remains unique within the entire dataset; *pcode* remains unique within the household.					

It is important for the data user to keep in mind that the CAPS was designed as a panel of young adults not a panel of households. All key individuals (young adults, older adults and children born to female young adults) are correctly and uniquely identified by *personid.* Other household members are mostly uniquely identified by *perrsonid* but there are instances where individuals have acquired more than one *personid* over the life of the panel. This can occur when non-key individuals move out and then later back in to households with young adults. For example, in Wave 1 a young adult lives with her mother and the mother is assigned a *pcode* of 2. In Wave 2 the young adult has moved out of home and lives with her boyfriend. In Wave 3 she has moved back to living with her mother. As she was not living with her mother when we last saw her we would view the mother as a new household member and assign them a *pcode* of 51. While efforts have been made to identify such cases, particularly in the case of parents, we advise users to proceed with caution when trying to identify non-key individuals over multiple waves.

9. Variable name conventions

This section contains information on the variable naming conventions used in the CAPS Wave 1-2-3-4 data.

Table 28 lists variable name abbreviations used (e.g. *birth month* is always represented by *bmth* in the variable name). For information regarding the variable prefix and suffix conventions see Table 2 in section 3.2 of *A Very Short Introduction to the Integrated Waves 1-2-3-4 (2002-2006) Data*.

Table 28: Variable abbreviation conventions

Abbreviation used	Full text
bmth	birth month
bst	best
byr	birth year
cal	calender
cd	code
ch	child
chg	change
chs	choose
comp	completed
cont	contact
contr	contraception
contrib	contribute
csg	child support grant
CT	cape town
cur	current
d	day
des	describe
disab	disabled
dth	death
ed	education
empl	employer/employed
exp	expenditure
expct	expect
fath	father
fin	financial
frnd	friend
ft	full time
grd	grade
grnt	grant
h	hour
hh	household
hlth	health
illn	illness
inc	income
infl	influence
intv	interview
jb	job
knw	know
loc	location
lst	last
lv	leave/live
lvl	level
mar	marriage/marital

mat	material
moth	mother
mov	move
mth	month
nhh	non household
nm	name
nr	non resident
num	number
oth	other (when not other specify)
p	partner
par	parent
pc	per capita
pd	paid
per	person
preg	pregnant
prob	problem
protct	protect/protection
prov	province
pt	part time
pup	pupil
qual	qualification
rel	relation
res	resident/residence
rslt	result
rsn	reason
sch	school
ser	serious
sibs	siblings
spnt	spent
srch	search
stp	stop
subj	subject
sup	support
sym	symbol
tch	teacher/teach
tert	tertiary
tot	total
trans	transport
wg	wage
wh	where
whn	when
wk	week
wnt	want
wrk	work
wrt	write
wtch	watch

y	why

Additional conventions

Other variables _o for numbered variables: _o1 _o2 for job table variables: _oj1 _oj2	
Ever variables varever where var could be preg mar matrc etc	
Year variables 01 02 …	2001 2002 etc
Job variables w#y_varname_j* (using variable rename) w#y_varnum_* (using variable number on the questionnaire) w#y_varname_* (using variable rename)	Mega job table variables original job variables original wave 2 job variables-wave 2a and 2b merged

10. Household income imputations

Values for missing Wave 1 and Wave 3 household income have been imputed and are used in the generation of per capita income and quintile variables. See Table 29 below for a breakdown of all Wave 1 and Wave 3 households by type of household income information. This information is contained in the variables *w1h_impute* (for Wave 1) and *w3h_impute* (for Wave 3) in the public data. Note that households reporting zero income also have imputed values. If preferred for analysis, these households, where *w1h_impute*=2 (Wave 1), can be restored to zero income.

In Wave 1, at 16%, households in predominantly white clusters have the highest proportion of missing income while households in Coloured areas have the second highest proportion (5%). African areas have the lowest proportion of households with missing income (2%), but the highest proportion and number of households with zero income (3%). Households in white clusters again have a high proportion of missing income in Wave 3 (18.13%), however this percentage is even higher for households in Coloured areas (21.79%).

Table 29. Origin of Wave 1 and Wave 3 household income information (*w1h_impute, w3h_impute*)

Origin of household income	Majority Population Group of Cluster			
	African	Coloured	White	Total
Wave 1				
0. Household income from w1h_d4 (*w1h_totinc1*)	85.66%	69.84%	49.17%	72.86%
	1,935	*1,422*	*472*	*3,829*
1. Household income from w1h_d5 (*w1h_totinc2*)	9.03%	24.17%	34.38%	19.52%
	204	*492*	*330*	*1,026*
2. Zero from w1h_d4 or w1h_d5, imputed	3.19%	0.79%	0.31%	1.73%
	72	*16*	*3*	*91*
3. Missing w1h_d4 and w1h_d5, imputed	2.12%	5.21%	16.15%	5.88%
	48	*106*	*155*	*309*
Total	100%	100%	100%	100%
	2,259	*2,036*	*960*	*5,255*
Wave 3				
0. Household income from w3h_d4 (*w1h_totinc1*)	74.1%	42.75%	44.71%	55.63%
	761	*510*	*148*	*1,419*
1. Household income from w3h_d5 (*w1h_totinc2*)	17.53%	35.04%	36.86%	28.22%
	180	*418*	*122*	*720*
2. Zero from w3h_d4 or w3h_d5, imputed	0.58%	0.42%	0.3	0.47%
	6	*5*	*1*	*12*
3. Missing w3h_d4 and w3h_d5, imputed	7.79%	21.79%	18.13%	15.68%
	80	*260*	*60*	*400*
Total	100%	100%	100%	100%
	1,027	*1,193*	*331*	*2,551*

Imputed values for both Wave 1 and Wave 3 household income were created by predicting log total household income in an ordinary least squares (OLS) regression using the following independent variables (omitted categories in parentheses): majority population group of cluster (African); total household members age 18+ (0); total household members under 18 (0); gender and population group of head of household (male, African); primary language spoken in the household (isiXhosa); five-year age groups for age of household head (15-19); head born in Cape Town (born outside of Cape Town); years of schooling of household head (0); post-school education of household head (none); respondent's classification of household's financial situation (very poor); anyone in household have a bank account (no); rent/own residence (rent); number of rooms in residence (1); flush toilet in residence (no); residence connected to electricity supply (no); paraffin used for lighting, heating or cooking (no); drinking water from piped internal source (no); anyone in household own: radio, television, video, landline telephone cellular telephone, refrigerator, stove, microwave oven, washing machine, bicycle, motorcycle, car, computer, more than five books (no); sub-place of residence (a formal African area).

Kernel densities of the effect of imputed values on the distribution of Wave 1 and Wave 3 household income and probit regressions for missing household income are included in Appendices A, B, C and D to this document. Imputed values are included in the creation of per capita income variables described below.

For households who reported total household income in brackets (*w1h/w3h_totinc2*) as opposed to a point estimate (*w1h/w3h_totinc1*), mean values of **totinc1* within the brackets were assigned. These values and the original point estimates were divided by household size

(*w1h/w3h_hhsize*) to create a measure of per capita income (*w1h/w3h_pcy*) for all households. Quintiles of per capita income (*w1h/ w3h_pcyquint*) were created without breaking groups at the same level of income, and as result the quintiles are not equal in size. The range of each quintile is included in the variable's value labels.

Similarly constructed variables for Wave 3 per capita income and quintiles are also included in the data. No household income information was collected in Wave 2.

11. Helpful hints for working with the CAPS Data

This section contains tips and useful commands for working with the CAPS data in Stata, as well as a recommended citation for CAPS.

11.1. Merging data

Household roster data (one record per household member) and household level data (one record per household) in any wave can be merged on the wave specific unique household identifier. For example, to merge the Wave 3 household roster and household data in Stata:

```
*open the roster data
use "\CAPSData\v0806\Public\capsw3.h.roster.v0806.dta", clear

*sort on unique household identifier
sort w3h_hhid

*merge on w3h_hhid with household level data
merge w3h_hhid using "\CAPSData\v0806\Public\capsw3.h.v0806.dta"
```

This Stata code will only work if the household level data is sorted by the field on which you want to merge (i.e. capsw3.h.v.dta must be sorted on w3h_hhid). If the data is sorted simply open it, sort it, save it as follows:

```
*open the household data
use "\CAPSData\v0806\Public\capsw3.h.v0806.dta", clear

*sort on unique household identifier
sort w3h_hhid

*save the data
save "\CAPSData\v0806\Public\capsw3.h.v0806.dta", replace
```

and then repeat the commands listed above.

Data at the level of the young adult can be found in a number of files. These files can be merged on *personid*. For example, to merge the Wave 1-2-3-4 young adult data with the life history calendar data and the literacy and numeracy module in Stata:

```
*open young adult data
use "\CAPSData\v0806\Public\capsw1234.y.v0806.dta", clear

*sort on personid
sort personid

*merge on personid with life history calendar data
merge personid using "\CAPSData\v0806\Public\capsw1.cal.wide.v0806.dta"

*drop the merge variable
drop _merge

*sort on personid
sort personid

*merge on personid with the literacy and numeracy module
merge personid using "\CAPSData\v0806\Public\capsw1.lne.v0806.dta"
```

The Wave 4 older adult data can be merged with the Wave 1 household roster on *personid*. In Stata,

```
*open older adult data
use "\CAPSData\v0806\Public\capsw4.o.v0806.dta", clear

*sort on personid
sort personid


*merge on personid with wave 1 roster
merge personid using "\CAPSData\v0806\Public\capsw1.h.roster.v0806.dta"

*keep only the older adults
drop if _m==2
```

11.2 Frequently asked questions about using the CAPS Waves 1-2-3-4 data

I don't understand the difference between the "long" and "wide" versions of the Young Adult calendars. Where do they come from?

The Wave 1 Young Adult Life History Calendar is comprised of the variables from the following questions in the Young Adult questionnaire: b1, b2a, b2b, b2c, b2d, b2e, b3, b4, b5, b6, b7, b8, b9, b10, b11, b12, e27, e28, e34, e35. These questions cover the topics of Household relocation, Living Arrangements, School and Pregnancy. These questions were asked for every year of the respondent's life from birth, up to and including the current year.

Please note that for questions b1-b4 a year is defined as the period between birthdays, such that the year of age 4 begins on the birthday at which the respondent turned 4 until the day before they turned 5. However, for questions b5-b12 and e27, e28, e34, e35, the year follows the calendar year. Therefore the year of age 4 corresponds to the year in which the respondent was age 4.

Because these questions were asked in this format, there are two possible ways to arrange the data, either "long form" or "wide form".

Long form: each year in the life of each respondent is its own record in the data. Therefore, there is more than one row in the data for each respondent. In this form, there is a variable "*w1y_calage*" that includes the corresponding age of each record.

Wide form: the data from the calendar is stored a format such that each respondent remains only 1 row in the data. Therefore, there is a variable for each question at each year. In this form, the variables are named in the format *variable_age*: *w1y_b1_3*, *w1y_b1_4*, *w1y_b1_5* etc.

Depending on how you want to work with the data, one or the other of these forms may be more suitable. In Stata, you can "reshape" the data from long to wide and vice versa using the "reshape" command. The reshape command asks you to specify ("i") the individual identifier(s) and ("j") a variable to identify the sub-observations in the data.

The appropriate "i" to use with the CAPS Wave 1 YA Life History Calendar is i(*personid*); this defines the respondents which in "long" form have multiple records-1 per year of their

lives. The appropriate "j" to use with the CAPS Life History Calendar is j(*w1y_calage*); this defines the age of each respondent at each row in the "long" form.

Similar the monthly calendar for waves 2, 3 and 4 are available in both “long” and “wide” form.

Is there a “reshaped long” version of the Wave 3 (2005) Residential and Schooling History”?

No, there is not a public release version of the school history data collected in Wave 3 reshaped “long” (with one record per YA per year)- the only reshaped data available are the Wave 1 YA Life History Calendar and the Waves 2-3 YA Monthly Calendar.

However, it is easy to reshape long the Residential and Schooling History variables on the unique individual identifier *personid.* Here is the STATA code:

```
#delimit cr;
/*applicable variables:  w3y_b8_* w3y_b11_* w3y_b12_* w3y_b12_o* w3y_b13_*
w3y_b13_o* */

/* reshape long: where “i” is personid, generating a new variable “year” to
distinguish between the multiple records generated per YA.*/
reshape long w3y_b6_ w3y_b8_ w3y_b11_ w3y_b12_ w3y_b12_o w3y_b13_
w3y_b13_o, i(personid) j(year);
```

I am confused about the biological mother & father variables in the household rosters (w1h_biomom, w1h_biodad, w3h_biomom, w3h_biodad, w4h_biodad, w4h_biomom). When I tabbed them in Stata, the range goes from 1-8 in addition to 94, 95, 98, 99. What do 1-8 represent?

The questions for these variables ask for the line number of the biological mother/father of each household member, if the parent is present in the household. Therefore, w1h_biomom==1 means that this household member's mother is line number 1 (*pcode*==1) in that household in Wave 1.

I am interested in using STATA’s survey (“svy”) commands with the CAPS data. What is the appropriate “svyset” command to describe the survey design?

The CAPS sample was stratified on population group of the Census enumeration area, which is captured in the variable *majpop*. In the data, the *cluster* variable contains a scrambled code for Census enumeration area. The appropriate weight variable will depend on the unit of analysis and whether you want to correct for both sample design and non-response. Please see sections 2 and 6 of this document for a discussion of the sample design and weights for the CAPS data.

Why can I not find the Module C Household roster variables for Waves 2a and 2b in the Waves 1-2-3-4 Young Adult data?

The Wave 2 household roster variables, even though they were asked as part of the Wave 2 Young Adult questionnaire, are included in a separate household roster file for Wave 2 (capsw2.h.roster.*v*.dta). These variables are therefore not included in the Waves 1-2-3-4 Young Adult data.

In the Wave 2a questionnaire, Module C is followed by Module E. Is there a Module D?

No, there is no Module D in the Wave 2a questionnaire.

When I look at the Waves 1 and 3 questionnaires, it seems that different codes have been used for education levels. Are these codes compatible?

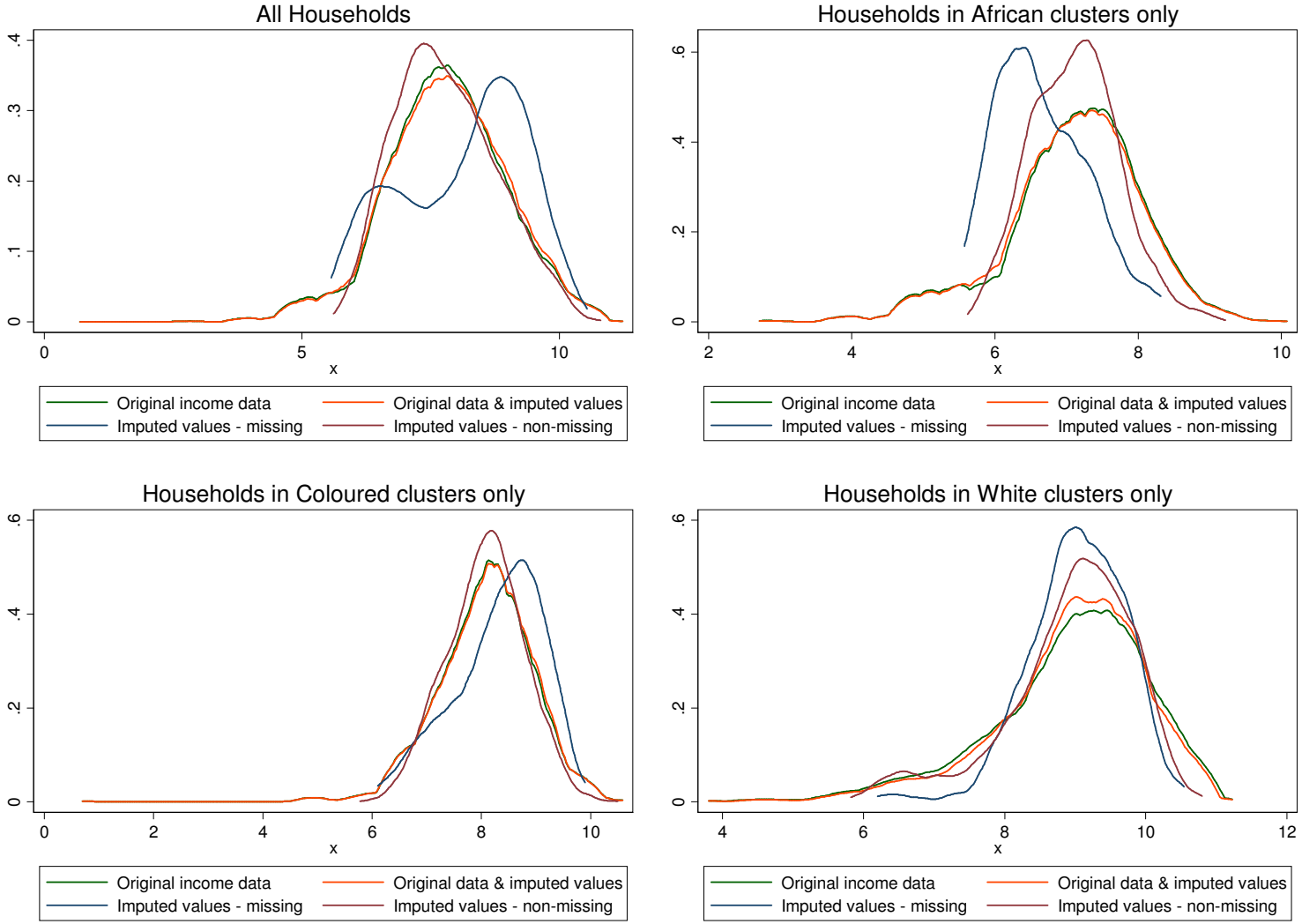
As you can see from the table below, we collapsed most of the answers in the range that was education codes 13-19 in Wave 1 in order to simplify the list. These new codes for Wave 3 were used in all education level questions. We used the new codes 26 and 27 in order to keep a clean separation of the two coding schemes when the waves are merged.

MATCH EDUCATION LEVEL CODES FROM WAVE 1 TO WAVE 3

WAVE 1	W1	W3	WAVE 3
Never enrolled/Grade Zero/Little Sub-A	0	95	No schooling/Grade 0/Little sub-A
Grade 1/Sub A	1	1	Grade 1/Sub A
Grade 2/Sub B	2	2	Grade 2/Sub B
Grade 3/Standard 1	3	3	Grade 3/Standard 1
Grade 4/Standard 2	4	4	Grade 4/Standard 2
Grade 5/Standard 3	5	5	Grade 5/Standard 3
Grade 6/Standard 4	6	6	Grade 6/Standard 4
Grade 7/Standard 5	7	7	Grade 7/Standard 5
Grade 8/Standard 6	8	8	Grade 8/Standard 6
Grade 9/Standard 7	9	9	Grade 9/Standard 7
Grade 10/Standard 8	10	10	Grade 10/Standard 8
Grade 11/Standard 9	11	11	Grade 11/Standard 9
Grade 12/Standard 10/Matric	12	12	Grade 12/Standard10/Matric
NTC I	**13**	**27**	Diploma/Cert that does not require matric, not from University or Technikon
NTC II	**14**	**27**	Diploma/Cert that does not require matric, not from University or Technikon
NTC III	**15**	**27**	Diploma/Cert that does not require matric, not from University or Technikon
Diploma/Cert with less than Grade 12/ Std 10 of **less** than 6 months duration	**16**	**27**	Diploma/Cert that does not require matric, not from University or Technikon
Diploma/Certificate with less than Grade 12/Std 10 of **more** than 6 months duration	**17**	**27**	Diploma/Cert that does not require matric, not from University or Technikon
Diploma/Cert **from an institution other than a Technikon/University** with Grade 12/Std 10 of **less than 6 months** duration	**18**	**26**	Diploma/Cert that requires matric, not from University or Techikon
Diploma/Cert **from an institution other than a Technikon/University** with Grade 12/Std 10 of **more than 6 months** duration	**19**	**26**	Diploma/Cert that requires matric, not from University or Techikon
Undergraduate Diploma/Certificate **from a Technikon** with Grade 12/Std 10*	20	20	Undergrad Diploma/Cert from a Technikon with Grade12/Std 10
Undergraduate Diploma **from a University** with Grade 12/Std 10*	21	21	Undergrad Diploma/Cert from a University with Grade12/Std 10
Undergraduate degree from a **Technikon**	22	22	Undergraduate degree from a Technikon
Undergraduate degree from a **University**	23	23	Undergraduate degree from a University
Postgraduate degree or diploma	24	24	Postgraduate degree or diploma
Other	25	25	Other (SPECIFY):
Don't know	99	99	Don't know

Appendices

Appendix A. Kernel Densities of Waves 1 Household Income – Effect of Imputed Values



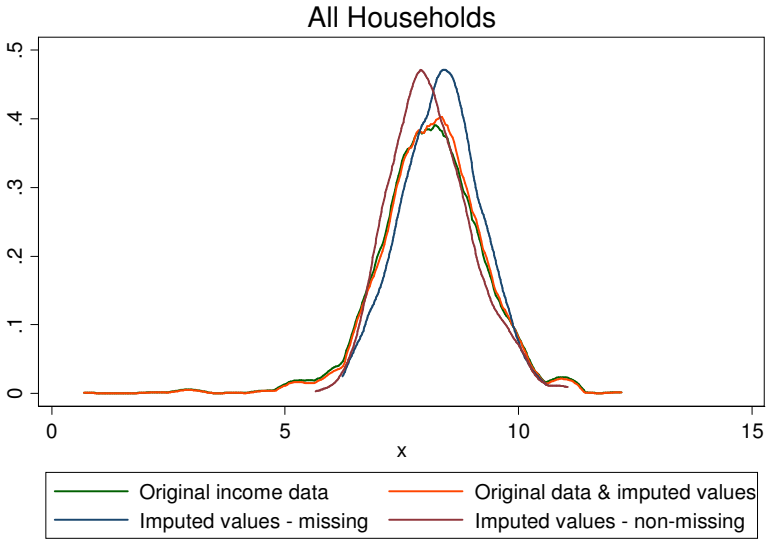
Appendix B. Probit regression for Wave 1 Households with missing income

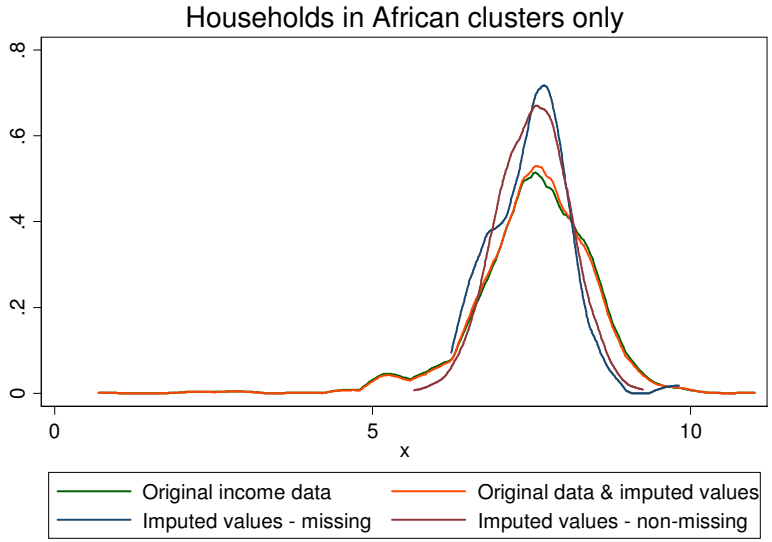
Variables	*Regression Coefficients and Standard Errors*		
	Coefficient	Standard Error	dF/dX
Cluster: majority Coloured	-0.185	[0.205]	-0.019
Cluster: majority White	0.164	[0.206]	0.019
Household members age 18+	-0.058	[0.027]**	-0.006
Household members age <18	-0.082	[0.025]***	-0.009
Head of household: female	-0.033	[0.061]	-0.003
Head of household: Coloured	0.561	[0.264]**	0.065
Head of household: White	0.502	[0.276]*	0.071
Household language: English	-0.232	[0.266]	-0.022
Household language: Afrikaans	-0.281	[0.265]	-0.028
Household language: other	-0.198	[0.295]	-0.018
Head of household: age 20-24	0.161	[0.308]	0.019
Head of household: age 25-29	0.014	[0.307]	0.001
Head of household: age 30-34	-0.194	[0.309]	-0.018
Head of household: age 35-39	-0.182	[0.305]	-0.017
Head of household: age 40-44	0.041	[0.302]	0.004
Head of household: age 45-49	-0.141	[0.306]	-0.014
Head of household: age 50-54	-0.022	[0.308]	-0.002
Head of household: age 55-59	0.081	[0.313]	0.009
Head of household: age 60-64	-0.455	[0.335]	-0.035
Head of household: age 65-69	-0.055	[0.330]	-0.006
Head of household: age 70-74	0.017	[0.340]	0.002
Head of household: age 75-79	0.074	[0.341]	0.008
Head of household: age missing	1.757	[0.438]***	0.498
Head of household: born in Cape Town	0.069	[0.070]	0.007
Head of household: years schooling 1-4	-0.326	[0.217]	-0.027
Head of household: years schooling 5-6	-0.045	[0.195]	-0.005
Head of household: years schooling 7	-0.097	[0.195]	-0.01
Head of household: years schooling 8	-0.095	[0.190]	-0.01
Head of household: years schooling 9	0.045	[0.196]	0.005
Head of household: years schooling 10	0.07	[0.189]	0.008
Head of household: years schooling 11	-0.026	[0.207]	-0.003
Head of household: years schooling 12	0.1	[0.191]	0.011
Head of household: years schooling missing	0.119	[0.223]	0.014
Head of household: Diploma/ certificate that does not require matric	-0.249	[0.222]	-0.022
Head of household: Diploma/ certificate that does require matric	0.046	[0.109]	0.005
Head of household: under/postgraduate degree from technikon or university	-0.264	[0.119]**	-0.023
Financially: very comfortable	-0.73	[0.196]***	-0.044
Financially: comfortable	-0.595	[0.120]***	-0.049
Financially: just getting by	-0.687	[0.106]***	-0.068
Financially: poor	-0.551	[0.101]***	-0.047

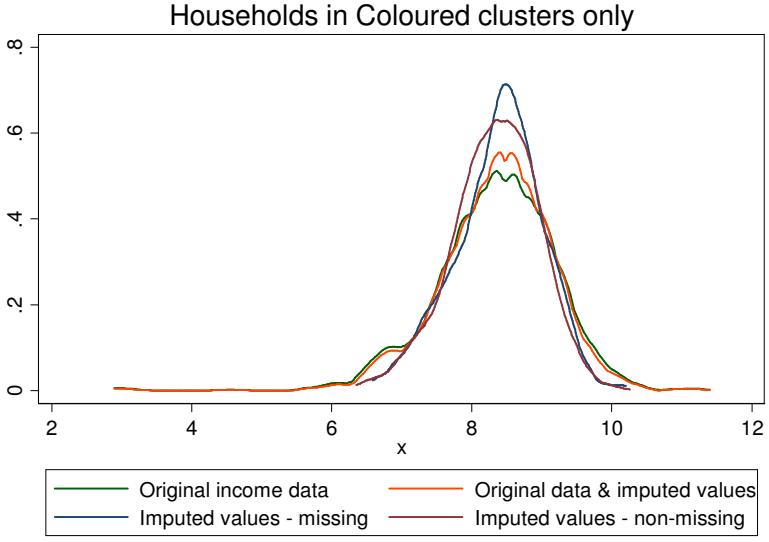
Financially: don't know	-0.245	[0.234]	-0.021
Someone has bank account	-0.289	[0.081]***	-0.034
Bank account: missing	0.404	[0.403]	0.059
Own residence	0.011	[0.073]	0.001
Own or rent residence: other	0.476	[0.460]	0.073
Number rooms in residence: 2	0.075	[0.125]	0.008
Number rooms in residence: 3	0.074	[0.134]	0.008
Number rooms in residence: 4	0.217	[0.129]*	0.025
Number rooms in residence: 5	0.379	[0.139]***	0.049
Number rooms in residence: 6	0.613	[0.156]***	0.097
Number rooms in residence: 7	0.584	[0.161]***	0.089
Number rooms in residence: don't know	0.669	[0.352]*	0.117
Residence has flush toilet	-0.16	[0.152]	-0.019
Residence has other type of toilet	-0.227	[0.169]	-0.021
Type of toilet: missing	0.541	[0.669]	0.087
Household uses paraffin	0.039	[0.120]	0.004
Connected to electricity supply	0.228	[0.134]*	0.021
Indoor piped water for drinking	-0.038	[0.112]	-0.004
own Radio, stereo	-0.059	[0.086]	-0.006
own Television	-0.255	[0.099]***	-0.03
own Video, VCR, DVD	0.062	[0.085]	0.007
own Telephone (not Cellular)	0.107	[0.089]	0.011
own Cellular telephone	0.035	[0.074]	0.004
own Refrigerator/freezer	-0.042	[0.111]	-0.005
own Gas/electric stove	-0.215	[0.104]**	-0.025
own Microwave	0.196	[0.096]**	0.021
own Washing Machine	0.075	[0.103]	0.008
own Bicycle	-0.044	[0.078]	-0.005
own Motorcycle	-0.279	[0.164]*	-0.024
own Car, Bakki or Kombi	0.192	[0.093]**	0.021
own Computer/laptop	0.001	[0.091]	0
own More than 5 books	-0.015	[0.082]	-0.002
Constant	-0.751	[0.391]*	
Observations	5255		
Pseudo R-squared	0.13		
Log likelihood	-1226.33		

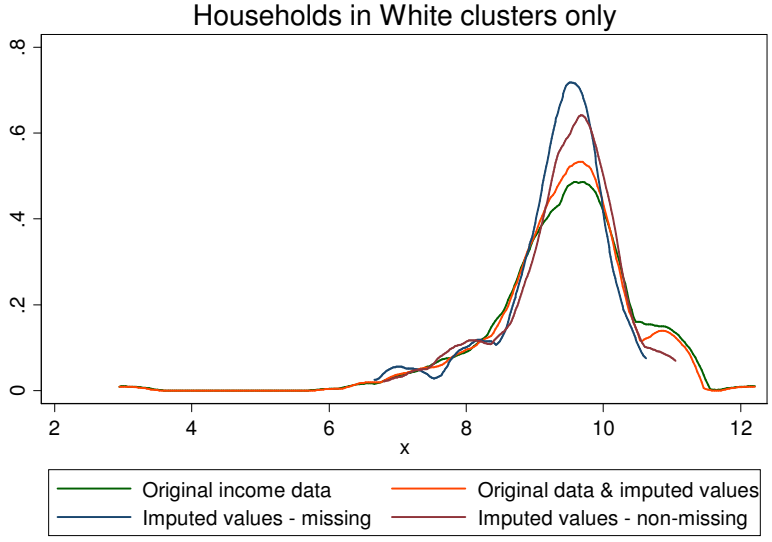
*Notes: Standard errors in brackets ; * significant at 10%; ** significant at 5%; *** significant at 1%*

Appendix C. Kernel Densities of Waves 3 Household Income – Effect of Imputed Values









Appendix D. Probit regression for Wave 3 Households with missing income

Variables	*Regression Coefficients and Standard Errors*		
	Coefficient	Standard Error	dF/dX
Cluster: majority Coloured	0.671	[0.280]**	0.151
Cluster: majority White	0.546	[0.284]*	0.146
Household members age 18+	0.086	[0.025]***	0.019
Household members age <18	-0.001	[0.025]	0
Head of household: female	0.034	[0.067]	0.007
Head of household: Coloured	0.525	[0.398]	0.116
Head of household: White	0.38	[0.420]	0.097
Household language: English	-0.436	[0.386]	-0.085
Household language: Afrikaans	-0.271	[0.385]	-0.057
Household language: other	-0.175	[0.427]	-0.035
Head of household: age 20-24	-0.9	[0.506]*	-0.121
Head of household: age 25-29	-0.644	[0.500]	-0.1
Head of household: age 30-34	-0.704	[0.514]	-0.104
Head of household: age 35-39	-0.416	[0.490]	-0.074
Head of household: age 40-44	-0.652	[0.483]	-0.109
Head of household: age 45-49	-0.445	[0.482]	-0.083
Head of household: age 50-54	-0.458	[0.484]	-0.084
Head of household: age 55-59	-0.481	[0.488]	-0.084
Head of household: age 60-64	-0.581	[0.497]	-0.094
Head of household: age 65-69	-0.414	[0.506]	-0.073
Head of household: age 70-74	-0.462	[0.524]	-0.078
Head of household: age 75-79	-0.457	[0.525]	-0.077
Head of household: age missing	-0.401	[0.489]	-0.072
Head of household: born in Cape Town	-0.156	[0.083]*	-0.034
Head of household: years schooling 1-4	-0.02	[0.282]	-0.004
Head of household: years schooling 5-6	-0.131	[0.274]	-0.027
Head of household: years schooling 7	-0.078	[0.268]	-0.016
Head of household: years schooling 8	0.108	[0.259]	0.025
Head of household: years schooling 9	0.168	[0.260]	0.04
Head of household: years schooling 10	0.079	[0.252]	0.018
Head of household: years schooling 11	0.121	[0.278]	0.028
Head of household: years schooling 12	0.238	[0.243]	0.056
Head of household: years schooling missing	0.086	[0.276]	0.02
Head of household: Diploma/ certificate that does not require matric	0.341	[0.479]	0.089
Head of household: under/postgraduate degree from technikon or university	0.112	[0.275]	0.026
Financially: very comfortable	-0.136	[0.198]	-0.028
Financially: comfortable	-0.031	[0.155]	-0.007
Financially: just getting by	-0.279	[0.149]*	-0.06
Financially: poor	-0.475	[0.179]***	-0.084
Someone has bank account	-0.238	[0.092]***	-0.056
Own residence	0.114	[0.080]	0.024
Number rooms in residence: 2	0.053	[0.176]	0.012

Number rooms in residence: 3	-0.128	[0.177]	-0.027
Number rooms in residence: 4	-0.255	[0.173]	-0.052
Number rooms in residence: 5	-0.314	[0.180]*	-0.062
Number rooms in residence: 6	-0.475	[0.193]**	-0.085
Number rooms in residence: 7	-0.342	[0.200]*	-0.064
Number rooms in residence: don't know	0.484	[0.545]	0.133
Residence has flush toilet	-0.01	[0.294]	-0.002
Residence has other type of toilet	-0.47	[0.316]	-0.08
Household uses paraffin	-0.025	[0.129]	-0.006
Household uses paraffin: don't know	0.819	[0.581]	0.254
Connected to electricity supply	0.011	[0.210]	0.002
Indoor piped water for drinking	-0.214	[0.120]*	-0.049
own Radio, stereo	0.03	[0.113]	0.006
own Television	-0.083	[0.139]	-0.019
own Video, VCR, DVD	0.055	[0.092]	0.012
own Telephone (not Cellular)	0.1	[0.086]	0.022
own Cellular telephone	0.21	[0.098]**	0.043
own Refrigerator/freezer	-0.232	[0.135]*	-0.055
own Gas/electric stove	-0.068	[0.129]	-0.015
own Microwave	0.143	[0.097]	0.031
own Washing Machine	0.073	[0.108]	0.016
own Bicycle	-0.109	[0.081]	-0.023
own Motorcycle	0.141	[0.173]	0.033
own Car, Bakki or Kombi	0.133	[0.088]	0.03
own Computer/laptop	-0.02	[0.098]	-0.004
own More than 5 books	-0.085	[0.089]	-0.019
Constant	-0.687	[0.649]	
Observations	2547		
Pseudo R-squared	0.09		
Log likelihood	-1023.12		

*Notes: Standard errors in brackets ; * significant at 10%; ** significant at 5%; *** significant at 1%*

References

Anderson, Kermyt G., Anne Case and David Lam. 2001. Causes and Consequences of Schooling Outcomes in South Africa: Evidence from Survey Data. *Social Dynamics* 27 (1), 37-59.

Barbarin, O., and Linda Richter. 2001. *Mandela's Children: Growing Up in Post-Apartheid South Africa*. London: Routledge.

Bray, Rachel. 2002. Missing Links? An Examination of Contributions Made by Social Surveys to our Understanding of Child Well-being in South Africa. *CSSR Working Paper no. 23*. Cape Town: Centre for Social Science Research, University of Cape Town.

Crankshaw, Owen, Matthew Welch, and Shirley Butcher (2001) "GIS Technology and Survey Sampling Methods: The Khayelitsha/Mitchell's Plain 2000 Survey, Social Dynamics, 27(2): 156-174.

May, Julian and Benjamin Roberts 2001. Panel Data and Policy Analysis in South Africa: Taking a Long View. *Social Dynamics* 27 (1): 96-119.

Russell, Margo. 2003a. Understanding Black Households: The Problem. *Social Dynamics* 28 (2): 5-47.

Russell, Margo. 2003b. Are Urban Black Families Nuclear? A Comparative Study of Black and White South African Family Norms. *Social Dynamics* 28 (2): 153-76.

Rutenberg N, Kehus-Alons C, Brown L, Macintyre K, Dallimore A, Kaufman C. (2001) *Transitions to Adulthood in the Context of AIDS in South Africa: Report of Wave I. Population Council, March 2001.*

Seekings, Jeremy. 2001. The Uneven Development of Quantitative Social Science in South Africa. *Social Dynamics* 27 (1), 1-36.

Shisana O, Rehle T, Simbayi LC, Parker W, Zuma K, Bhana A, Connolly C, Jooste S, Pillay V et al. (2005) *South African National HIV Prevalence, HIV Incidence, Behaviour and Communication Survey, 2005*. Cape Town: HSRC Press

South Africa Labour Development Research Unit. (1994) Project for Statistics on Living Standards and Development (PSLSD) *South Africans Rich and Poor: Baseline Household Statistics***.** University of Cape Town, South Africa.